

8/453
31-8-76

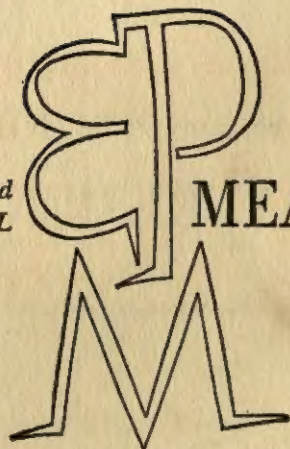
EDUCATIONAL AND PSYCHOLOGICAL
MEASUREMENT

Volume XXV

1965

BOX 6907, COLLEGE STATION, DURHAM N. C.

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Editor: G. FREDERIC KUDER

Associate Editor: W. Scott Gehman

Managing Manager: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

WILLIAM V. CLEMANS

Science Research Associates, Inc.

LOUIS D. COHEN

University of Florida

HAROLD A. EDGERTON

Performance Research, Incorporated

MAX D. ENGELHART

Chicago City Junior Colleges

E. B. GREENE

Chrysler Corporation

J. P. GUILFORD

University of Southern California

JOHN A. HORNADAY

Houghton Mifflin Company

E. F. LINDQUIST

State University of Iowa

FREDERIC M. LORD

Educational Testing Service

ARDIE LUBIN

U. S. Naval Hospital, San Diego

SAMUEL MESSICK

Educational Testing Service

WILLIAM B. MICHAEL

*University of California
Santa Barbara*

HOWARD G. MILLER

*North Carolina State University
at Raleigh*

P. J. RULON

Harvard University

C. L. SHARTLE

Ohio State University

KENDON SMITH

*The University of North Carolina
at Greensboro*

THELMA G. THURSTONE

*University of North Carolina
at Chapel Hill*

HERBERT A. TOOPS

Ohio State University

JOHN E. WILLIAMS

Wake Forest College

E. G. WILLIAMSON

University of Minnesota

DOROTHY ADKINS WOOD

*University of North Carolina
at Chapel Hill*

INDEX FOR VOLUME XXV

AGER, JOEL (WITH DORIS ALLEN). <i>A Factor Analytic Study of the Ability to Spell</i>	153
AIKEN, LEWIS R., JR. <i>The Probability of Chance Success on Objective Test Items</i>	127
ALLEN, DORIS (WITH JOEL AGER). <i>A Factor Analytic Study of the Ability to Spell</i>	153
ANZEL, ANNE SMITH (WITH DANIEL V. CAPUTO, JON M. PLAPP, AND CONSTANCE HANF). <i>The Validity of the Edwards Personal Preference Schedule (EPPS) Employing Projective and Behavioral Criteria</i>	829
BAKER, FRANK B. (WITH THOMAS J. MARTIN). <i>An IPL-V Technique for Simulation Programs</i>	859
BALINKY, JEAN L. <i>The Application of a Configuration Method to the Prediction of Success in First Grade</i>	605
BARGER, BEN (WITH EVERETTE HALL). <i>The Interaction of Ability Levels and Socioeconomic Variables in the Prediction of College Dropouts and Grade Achievement</i>	501
BARGER, BEN (WITH JAY L. CHAMBERS AND LEWIS R. LIEBERMAN). <i>Need Patterns and Abilities of College Dropouts</i>	509
BASHAW, W. L. <i>A FORTRAN Program for a Central Prediction System</i>	201
BERKSHIRE, JAMES R. (WITH ROBERT J. WHERRY, JR. AND ROBERT W. SHOENBERGER). <i>Secondary Selection in Naval Aviation Training</i>	191
BISHOP, CAROL H. (WITH JASON MILLMAN AND ROBERT EBEL). <i>An Analysis of Test-Wiseness</i>	707
BLACK, MICHAEL S. <i>The Development of Personality Factors in Children and Adolescents</i>	767
BLUMENFELD, WARREN S. <i>Predicting Grade Point Average with the SRA Tests of Educational Ability: A 13-Month Follow-Up Study</i>	555
BOYCE, RICHARD W. (WITH R. C. PAXSON). <i>The Predictive Validity of Eleven Tests at One State College</i>	1143
BOYER, ERNEST L. (WITH WILLIAM B. MICHAEL). <i>Uses of Cognitive and Non-Cognitive Test Measures in Sixty-four Private Liberal Arts Colleges: Implications for Predictive Validity and Assessment of Change</i>	1157
BROWN, STEPHEN W. (WITH WILLIAM B. MICHAEL AND RUSSELL HANEY). <i>The Predictive Validity of a Battery of Diversified Measures Relative to Success in Student Nursing</i>	579
CAPUTO, DANIEL V. (WITH CONSTANCE HANF). <i>The EPPS Pattern and the "Nursing Personality."</i>	421

CAPUTO, DANIEL V. (WITH JON M. PLAPP AND GEORGE PSATHAS). <i>Intellective Predictors of Success in Nursing School</i>	565
CAPUTO, DANIEL V. (WITH JON M. PLAPP, CONSTANCE HANF, AND ANNE SMITH ANZEL). <i>The Validity of the Edwards Personal Preference Schedule (EPPS) Employing Projective and Behavioral Criteria</i>	829
CARSE, WILLIAM T. (WITH ERNEST D. MCDANIEL). <i>Validation of the Kahn Intelligence Tests</i>	1152
CHAMBERS, JAY L. (WITH BEN BARGER AND LEWIS R. LIEBERMAN). <i>Need Patterns and Abilities of College Dropouts</i>	509
CHANSKY, NORMAN M. <i>Aptitude, Personality, and Achievement in Six College Curricula</i>	1117
CONRY, ROBERT (WITH WALTER T. PLANT). <i>WAIS and Group Test Predictions of an Academic Success Criterion: High School and College</i>	493
COURSON, CLIFFORD C. <i>The Use of Inference as a Research Tool</i>	1029
CRAWFORD, PAUL L. <i>An Evaluation of an Attitude Scale toward Teaching</i>	535
CRONBACH, LEE J. (WITH PETER SCHONEMANN AND DOUGLAS MCKIE). <i>Alpha Coefficients for Stratified-Parallel Tests</i>	291
CURETON, EDWARD E. <i>Reliability and Validity: Basic Assumptions and Experimental Designs</i>	327
DATTEL, WILLIAM E. (WITH FORREST D. HALL AND CHARLES P. RUFFE). <i>Measurement of Achievement Motivation in Army Security Agency Foreign Language Candidates</i>	539
DEMPSEY, PAUL. <i>The Concept of Item Difficulty in Personality Measurement: A Note on Hanley's Formulation</i>	143
DE SENA, PAUL A. (WITH LOUISE ANN WEBER). <i>The Predictive Validity of the School College Ability Test (SCAT) and the American College Test (ACT) at a Liberal Arts College for Women</i>	1149
DICK, WALTER (WITH RICHARD E. SPENCER). <i>An Application of Computer Programing to Test Analysis and Item Analysis</i>	211
DIESENHAUS, HERMAN (WITH KENNETH I. HOWARD). <i>16 PF Item Response Patterns as a Function of Repeated Testing</i>	365
DIGMAN, JOHN M. <i>Child Behavior Ratings: Further Evidence of a Multiple-Factor Model of Child Personality</i>	787
DIZNEY, HENRY F. (WITH KAORU YAMAOTO). <i>Effects of Three Sets of Test Instructions on Scores on an Intelligence Scale</i>	87
DIZNEY, HENRY F. (WITH ELINOR A. ELFNER AND HORACE A. PAGE). <i>American College Test (ACT) Performance as a</i>	

<i>Function of Examinee Acceptance of Test</i>	547
DIZNEY, HENRY. <i>Concurrent Validity of the Test of English as a Foreign Language for a Group of Foreign Students at an American University</i>	1129
DOMINO, GEORGE. <i>A Validation of Howard's Test of Change-Seeking Behavior</i>	1073
EBEL, ROBERT (WITH JASON MILLMAN AND CAROL H. BISHOP). <i>An Analysis of Test-Wiseness</i>	707
EDWARDS, DOROTHY S. <i>Test Scoring and Analysis with the Film-Optical Sensing Device for Input to Computers</i>	221
ELFNER, ELINOR A. (WITH HENRY F. DIZNEY AND HORACE A. PAGE). <i>American College Test (ACT) Performance as a Function of Examinee Acceptance of Test</i>	547
ELLIOTT, JAMES M. (WITH H. G. OSBURN). <i>The Effects of Partial-Pacing on Test Parameters</i>	347
EMERSON, PHILLIP L. <i>A FORTRAN Generator of Polynomials Orthonormal over Unequally Spaced and Weighted Ab-scissas</i>	867
FEDER, DANIEL D. <i>Intriguing Problems of Design in Pre-dicting College Success</i>	29
FORD, LEROY H., JR. (WITH MURRAY MEISELS). <i>Social Desir-ability and the Semantic Differential</i>	465
FRENCH, JOHN W. <i>The Relationship of Problem-Solving Styles to the Factor Composition of Tests</i>	9
GAMES, PAUL A. <i>Scorit-A FORTRAN Program for Scoring and Item Analysis of Porta-Punch Test Cards</i>	881
GOLDBERG, LEWIS R. (WITH JERRY S. WIGGINS). <i>Interrelation-ships among MMPI Item Characteristics</i>	381
GOLDBERG, LEWIS R. (WITH LEONARD G. RORER). <i>Acquiescence in the MMPI?</i>	801
GOLDMAN, IRWIN J. <i>Acceptance of Sc Scale Statements by Visual Art Students</i>	819
GRAHAM, WILLIAM K. (WITH S. S. KOMORITA). <i>Number of Scale Points and the Reliability of Scales</i>	987
GRIMSLEY, GLEN (WITH GEORGE W. SUMMERS). <i>Selection Tech-niques for Pakistani Postgraduate Students of Business</i>	1133
GUILFORD, J. P. <i>The Minimal Phi Coefficient and the Maxi-mal Phi</i>	3
GUILFORD, J. P. (WITH RALPH HOEPFNER AND HUGH PETER-son). <i>Predicting Achievement in Ninth-Grade Mathe-matics from Measures of Intellectual-Aptitude Factors</i> ..	659
GUTHRIE, GEORGE M. (WITH WILLARD E. REITZ AND PHILLIP S. VERY). <i>Experience, Expertness, and Ideal Teaching Re-lationships</i>	1051
GUTTMAN, ISAIAH (WITH NAMBURY S. RAJU). <i>Correlation as a Function of Predictor Score Points</i>	655

GUTTMAN, ISAIAH (WITH NAMBURY S. RAJU). <i>A New Working Formula for the Split-Half Reliability Model</i>	963
HALL, EVERETTE (WITH BEN BARGER). <i>The Interaction of Ability Levels and Socioeconomic Variables in the Prediction of College Dropouts and Grade Achievement</i>	501
HALL, FORREST D. (WITH WILLIAM E. DATEL AND CHARLES P. RUFÉ). <i>Measurement of Achievement Motivation in Army Security Agency Foreign Language Candidates</i> ...	539
HANEY, RUSSELL (WITH WILLIAM B. MICHAEL AND STEPHEN W. BROWN). <i>The Predictive Validity of a Battery of Diversified Measures Relative to Success in Student Nursing</i>	579
HANF, CONSTANCE (WITH DANIEL V. CAPUTO). <i>The EPPS Pattern and the "Nursing Personality"</i>	421
HANF, CONSTANCE (WITH DANIEL V. CAPUTO, JON M. PLAPP, AND ANNE SMITH ANZEL). <i>The Validity of the Edwards Personal Preference Schedule (EPPS) Employing Projective and Behavioral Criteria</i>	829
HARTNETT, RODNEY T. (WITH CLIFFORD T. STEWART). <i>Personality Rigidity of Students Showing Consistent Discrepancies between Instructor Grades and Term-End Examination Grades</i>	1111
HEATH, HELEN A. <i>Time-Saving Procedure for Computing Z Scores</i>	323
HEILBRUN, ALFRED B., JR. <i>The Social Desirability Variable: Implications for Test Reliability and Validity</i>	745
HIMELSTEIN, PHILIP. <i>Validities and Intercorrelations of MMPI Subscales Predictive of College Achievement</i>	1125
HOEPFNER, RALPH (WITH J. P. GUILFORD AND HUGH PETERSON). <i>Predicting Achievement in Ninth-Grade Mathematics from Measures of Intellectual-Aptitude Factors</i>	659
HOPKINS, KENNETH D. (WITH CAROLYN J. WILKERSON). <i>Differential Content Validity: The California Spelling Test, an Illustrative Example</i>	413
HORN, JOHN L. <i>An Empirical Comparison of Methods for Estimating Factor Scores</i>	313
HOWARD, KENNETH I. (WITH HERMAN DISENHAUS). <i>16 PF Item Response Patterns as a Function of Repeated Testing</i>	365
HOWAT, M. GORDON. <i>Variables Affecting the Graduate Assistant in a Computer Training Position</i>	887
HUGHES, HERBERT H. (WITH W. EUGENE TRIMBLE). <i>The Use of Complex Alternatives in Multiple Choice Items</i>	117
IVEY, ALLEN E. (WITH MARK B. PETERSON). <i>Vocational Preference Patterns of Communications Graduates</i>	849

IZARD, CARROLL E. (WITH JUM C. NUNNALLY). <i>Evaluative Responses to Affectively Positive and Negative Facial Photographs: Factor Structure and Construct Validity</i> ..	1061
JACOBSON, MILTON D. <i>Reading Difficulty of Physics and Chemistry Textbooks</i>	449
JASPEN, NATHAN. <i>Polyserial Correlation Programs in FORTRAN</i>	229
JASPEN, NATHAN. <i>A Subroutine to Refine the Inverse of a Matrix</i>	873
JASPEN, NATHAN. <i>The Calculation of Probabilities Corresponding to Values of z, t, F, and Chi-Square</i>	877
JOHNSON, M. CLEMENS. <i>Computer Search for Group Differences</i>	239
JONES, CHARLES W. (WITH DAN McMILLEN). <i>Engineering Freshman Norms for the D.A.T. Mechanical Reasoning and Space Relations Tests Utilizing Fifteen-Minute Time Limits</i>	459
JONES, ROBERT A. (WITH CALVIN PULLIAS AND WILLIAM B. MICHAEL). <i>An IBM 1401 Computer Program for Item and Test Analysis</i>	217
KASPAR, JOSEPH C. (WITH FRANCES M. THRONE AND JEROME L. SCHULMAN). <i>The Peabody Picture Vocabulary Test in Comparison with Other Intelligence Tests and an Achievement Test in a Group of Mentally Retarded Boys</i>	589
KIPNIS, DAVID. <i>The Relationship between Persistence, Insolence, and Performance, as a Function of General Ability</i>	95
KIPNIS, DAVID (WITH CARL WAGNER). <i>The Interaction of Personality and Intelligence in Task Performance</i>	731
KLUGH, HENRY E. (WITH ROBERT D. TARTE). <i>Alternations after Forced Choices as a Function of Dominance in Women</i>	149
KOMORITA, S. S. (WITH WILLIAM K. GRAHAM). <i>Number of Scale Points and the Reliability of Scales</i>	987
LEWIS, JOHN W. (WITH ARTHUR MITTMAN). <i>Correlates of Achievement on the Admissions Tests for Graduate Study in Business</i>	585
LIEBERMAN, LEWIS R. (WITH JAY L. CHAMBERS AND BEN BARGER). <i>Need Patterns and Abilities of College Dropouts</i>	509
LOVELL, VICTOR R. (WITH JERRY S. WIGGINS). <i>Communality and Favorability as Sources of Method Variance in the MMPI</i>	399
MADAUS, GEORGE F. (WITH JOHN J. WALSH). <i>Departmental Differentials in the Predictive Validity of the Graduate Record Examination Aptitude Tests</i>	1105
MARKS, EDMOND (WITH JOSEPH E. MURRAY). <i>Nonadditive</i>	

<i>Effects in the Prediction of Academic Achievement</i>	1097
MARTIN, THOMAS J. (WITH FRANK B. BAKER). <i>An IPL-V Technique for Simulation Programs</i>	859
MATTSON, DALE. <i>The Effects of Guessing on the Standard Error of Measurement and the Reliability of Test Scores</i>	727
MATTSON, DALE E. <i>A Generalization of the Median Test</i>	1023
MAY, FRANK B. (WITH ALAN W. METCALF). <i>A Factor-Analytic Study of Spontaneous-Flexibility Measures</i>	1039
MAYHEW, LEWIS B. <i>Non-Test Predictors of Academic Achievement</i>	39
MCDANIEL, ERNEST D. (WITH WILLIAM T. CARSE). <i>Validation of the Kahn Intelligence Tests</i>	1152
McKIE, DOUGLAS (WITH LEE J. CRONBACH AND PETER SCHONEMANN). <i>Alpha Coefficients for Stratified-Parallel Tests</i>	291
McMILLEN, DAN (WITH CHARLES W. JONES). <i>Engineering Freshman Norms for the D.A.T. Mechanical Reasoning and Space Relations Tests Utilizing Fifteen-Minute Time Limits</i>	459
McQUITTY, LOUIS L. <i>A Conjunction of Rank Order Typal Analysis and Item Selection</i>	949
MEISELS, MURRAY (WITH LEROY H. FORD, JR.). <i>Social Desirability and the Semantic Differential</i>	465
METCALF, ALAN W. (WITH FRANK B. MAY). <i>A Factor-Analytic Study of Spontaneous-Flexibility Measures</i>	1039
MEYERS, C. E. (WITH PHILLIP WEISE AND JOHN K. TUEL). <i>PMA Factors, Sex, and Teacher Nomination in Screening Kindergarten Gifted</i>	597
MICHAEL, WILLIAM B. <i>Measurement and Prediction in the College Admissions Process: Some Possible Directions for Future Research</i>	55
MICHAEL, WILLIAM B. (WITH ROBERT A. JONES AND CALVIN PULLIAS). <i>An IBM 1401 Computer Program for Item and Test Analysis</i>	217
MICHAEL, WILLIAM B. (WITH RUSSELL HANEY AND STEPHEN W. BROWN). <i>The Predictive Validity of a Battery of Diversified Measures Relative to Success in Student Nursing</i>	579
MICHAEL, WILLIAM B. (WITH ERNEST L. BOYER). <i>Uses of Cognitive and Non-Cognitive Test Measures in Sixty-four Private Liberal Arts Colleges: Implications for Predictive Validity and Assessment of Change</i>	1157
MILLMAN, JASON (WITH CAROL H. BISHOP AND ROBERT EBEL). <i>An Analysis of Test-Wiseness</i>	707
MITTMAN, ARTHUR (WITH JOHN W. LEWIS). <i>Correlates of Achievement on the Admissions Test for Graduate Study</i>	

<i>in Business</i>	585
MUKHERJEE, BISHWA NATH. <i>The Prediction of Grades in Introductory Psychology from Tests of Primary Mental Abilities</i>	557
MULLINS, CECIL J. (WITH H. G. OSBURN AND DANIEL E. SHEER). <i>Validation of a Carefulness Test Battery</i>	525
MURRAY, JOSEPH E. (WITH EDMOND MARKS). <i>Nonadditive Effects in the Prediction of Academic Achievement</i>	1097
MYERS, ALBERT E. <i>Risk Taking and Academic Success and Their Relation to an Objective Measure of Achievement Motivation</i>	355
NAYLOR, JAMES C. (WITH ROBERT J. WHERRY, SR.). <i>The Use of Simulated Stimuli and the "JAN" Technique to Capture and Cluster the Policies of Raters</i>	969
NICHOLS, ROBERT C. (WITH WILLIAM TETZLAFF). <i>Test Scoring and Item Analysis Programs</i>	205
NUNNALLY, JUM C. (WITH CARROLL E. IZARD). <i>Evaluative Responses to Affectively Positive and Negative Facial Photographs: Factor Structure and Construct Validity</i> ..	1061
OHNMACHT, FRED W. <i>Factor Analysis of Ranked Educational Objectives: An Approach to Value Orientation</i>	437
OSBURN, H. G. (WITH JAMES M. ELLIOTT). <i>The Effects of Partial-Pacing on Test Parameters</i>	347
OSBURN, H. G. (WITH CECIL J. MULLINS AND DANIEL E. SHEER). <i>Validation of a Carefulness Test Battery</i>	525
OVERALL, JOHN E. <i>Reliability of Composite Ratings</i>	1011
PAGE, HORACE A. (WITH HENRY F. DIZNEY AND ELINOR A. ELFNER). <i>American College Test (ACT) Performance as a Function of Examinee Acceptance of Test</i>	547
PAXSON, R. C. (WITH RICHARD W. BOYCE). <i>The Predictive Validity of Eleven Tests at One State College</i>	1143
PERLOFF, ROBERT (WITH LEROY WOLINS). <i>The Factorial Composition of AGCT "Subtests" along with College Aptitude Items and High School Grades</i>	73
PERLOFF, ROBERT (WITH LEROY WOLINS). <i>Item Difficulty as a Function of Perceived Item Directions</i>	79
PETERSON, HUGH (WITH J. P. GUILFORD AND RALPH HOEFFNER). <i>Predicting Achievement in Ninth-Grade Mathematics from Measures of Intellectual-Aptitude Factors</i>	659
PETERSON, MARK B. (WITH ALLEN E. IVEY). <i>Vocational Preference Patterns of Communications Graduates</i>	849
PLANT, WALTER T. (WITH ROBERT CONRY). <i>WAIS and Group Test Predictions of an Academic Success Criterion: High School and College</i>	493
PLAPP, JON M. (WITH GEORGE PSATHAS AND DANIEL V. CAPUTO). <i>Intellective Predictors of Success in Nursing</i>	

<i>School</i>	565
PLAPP, JON M. (WITH DANIEL V. CAPUTO, CONSTANCE HANF, AND ANNE SMITH ANZEL). <i>The Validity of the Edwards Personal Preference Schedule (EPPS) Employing Projective and Behavioral Criteria</i>	829
PRESTON, RALPH C. <i>The Multiple-Choice Test as an Instrument in Perpetuating False Concepts</i>	111
PSATHAS, GEORGE (WITH JON M. PLAPP AND DANIEL V. CAPUTO). <i>Intellective Predictors of Success in Nursing School</i>	565
PULLIAS, CALVIN (WITH ROBERT A. JONES AND WILLIAM B. MICHAEL). <i>An IBM 1401 Computer Program for Item and Test Analysis</i>	217
RAJU, NAMBURY S. (WITH ISAIAH GUTTMAN). <i>Correlation as a Function of Predictor Score Points</i>	655
RAJU, NAMBURY S. (WITH ISAIAH GUTTMAN). <i>A New Working Formula for the Split-Half Reliability Model</i>	963
REGAN, MARY C. <i>Development and Classification of Models for Multivariate Analysis</i>	997
REITZ, WILLARD E. (WITH PHILLIP S. VERY AND GEORGE M. GUTHRIE). <i>Experience, Expertness, and Ideal Teaching Relationships</i>	1051
RIPPEY, ROBERT M. <i>A 1620 FORTRAN Program for Compiling a Flanders-Amidon Interaction Analysis Matrix</i>	235
RORER, LEONARD G. (WITH LEWIS R. GOLDBERG). <i>Acquiescence in the MMPI?</i>	801
ROSS, JOHN (WITH LAWRENCE J. STRICKER AND HAROLD SCHIFFMAN). <i>Prediction of College Performance with the Myers-Briggs Type Indicator</i>	1081
RUFE, CHARLES P. (WITH WILLIAM E. DATEL AND FORREST D. HALL). <i>Measurement of Achievement Motivation in Army Security Agency Foreign Language Candidates</i> ...	539
SCHIFFMAN, HAROLD (WITH LAWRENCE J. STRICKER AND JOHN ROSS). <i>Prediction of College Performance with the Myers-Briggs Type Indicator</i>	1081
SCHONEMANN, PETER (WITH LEE J. CRONBACH AND DOUGLAS MCKIE). <i>Alpha Coefficients for Stratified-Parallel Tests</i>	291
SCHULMAN, JEROME L. (WITH FRANCES M. THRONE AND JOSEPH C. KASPAR). <i>The Peabody Picture Vocabulary Test in Comparison with Other Intelligence Tests and an Achievement Test in a Group of Mentally Retarded Boys</i>	589
SCOTT, CARRIE M. <i>The Predictive Value of a Beginning First-Grade Intelligence Examination</i>	613
SHEER, DANIEL E. (WITH H. G. OSBURN AND CECIL J. MUL-LINS). <i>Validation of a Carefulness Test Battery</i>	525
SHOENBERGER, ROBERT W. (WITH JAMES R. BERKSHIRE AND	

ROBERT J. WHERRY, JR.). <i>Secondary Selection in Naval Aviation Training</i>	191
SINGER, HARRY. <i>Validity of the Durrell-Sullivan Reading Capacity Test</i>	479
SMITH, CHARLES P. <i>The Influence on Test Anxiety Scores of Stressful versus Neutral Conditions of Test Administration</i>	135
SPENCER, RICHARD E. (WITH WALTER DICK). <i>An Application of Computer Programing to Test Analysis and Item Analysis</i>	211
STEININGER, MARION. <i>Situational and Individual Determinants of Attitude Scale Responses</i>	757
STEWART, CLIFFORD T. (WITH RODNEY T. HARTNETT). <i>Personality Rigidity of Students Showing Consistent Discrepancies between Instructor Grades and Term-End Examination Grades</i>	1111
STODOLA, QUENTIN C. <i>Data Processing Procedure to Improve Classroom Testing</i>	885
STRICKER, LAWRENCE J. <i>Difficulty and Other Correlates of Criticalness Response Style at the Item Level</i>	683
STRICKER, LAWRENCE J. (WITH HAROLD SCHIFFMAN AND JOHN ROSS). <i>Prediction of College Performance with the Myers-Briggs Type Indicator</i>	1081
SUMMERS, GEORGE W. (WITH GLEN GRIMSLEY). <i>Selection Techniques for Pakistani Postgraduate Students of Business</i>	1133
TARTE, ROBERT D. (WITH HENRY E. KLUGH). <i>Alternations After Forced Choices as a Function of Dominance in Women</i>	149
TETZLAFF, WILLIAM (WITH ROBERT C. NICHOLS). <i>Test Scoring and Item Analysis Programs</i>	205
THOMAS, R. MURRAY. <i>A Rationale for Measurement in the Visual Arts</i>	163
THRONE, FRANCES M. (WITH JOSEPH C. KASPAR AND JEROME L. SCHULMAN). <i>The Peabody Picture Vocabulary Test in Comparison with Other Intelligence Tests and an Achievement Test in a Group of Mentally Retarded Boys</i>	589
TRIMBLE, W. EUGENE (WITH HERBERT H. HUGHES). <i>The Use of Complex Alternatives in Multiple Choice Items</i>	117
TUEL, JOHN K. (WITH PHILLIP WEISE AND C. E. MEYERS). <i>PMA Factors, Sex, and Teacher Nomination in Screening Kindergarten Gifted</i>	597
VERY, PHILLIP S. (WITH WILLARD E. REITZ AND GEORGE M. GUTHRIE). <i>Experience, Expertness, and Ideal Teaching Relationships</i>	1051
WAGNER, CARL (WITH DAVID KIPNIS). <i>The Interaction of</i>	

<i>Personality and Intelligence in Task Performance</i>	731
WALSH, JOHN J. (WITH GEORGE F. MADAUS). <i>Departmental Differentials in the Predictive Validity of the Graduate Record Examination Aptitude Tests</i>	1105
WALTON, WESLEY W. <i>Potentialities of the Computer for Measurement and Prediction with Respect to the College Admissions Process</i>	47
WEBB, SAM C. <i>Two Cross Validations of the Opinion, Attitude and Interest Survey</i>	517
WEBER, LOUISE ANN (WITH PAUL A. DE SENA). <i>The Predictive Validity of the School College Ability Test (SCAT) and the American College Test (ACT) at a Liberal Arts College for Women</i>	1149
WEISE, PHILLIP (WITH C. E. MEYERS AND JOHN K. TUEL). <i>PMA Factors, Sex, and Teacher Nomination in Screening Kindergarten Gifted</i>	597
WHERRY, ROBERT J., JR. (WITH JAMES R. BERKSHIRE AND ROBERT W. SHOENBERGER). <i>Secondary Selection in Naval Aviation Training</i>	191
WHERRY, ROBERT J., SR. (WITH JAMES C. NAYLOR). <i>The Use of Simulated Stimuli and the "JAN" Technique to Capture and Cluster the Policies of Raters</i>	969
WIGGINS, JERRY S. (WITH LEWIS R. GOLDBERG). <i>Interrelationships among MMPI Item Characteristics</i>	381
WIGGINS, JERRY S. (WITH VICTOR R. LOVELL). <i>Communalities and Favorability as Sources of Method Variance in the MMPI</i>	399
WILKERSON, CAROLYN J. (WITH KENNETH D. HOPKINS). <i>Differential Content Validity: The California Spelling Test, an Illustrative Example</i>	413
WOLINS, LEROY (WITH ROBERT PERLOFF). <i>The Factorial Composition of AGCT "Subtests" along with College Aptitude Items and High School Grades</i>	73
WOLINS, LEROY (WITH ROBERT PERLOFF). <i>Item Difficulty as a Function of Perceived Item Directions</i>	79
YAMAMOTO, KAORU (WITH HENRY F. DIZNEY). <i>Effects of Three Sets of Test Instructions on Scores on an Intelligence Scale</i>	87

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Editor: G. Frederic Kuder

Associate Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

WILLIAM V. CLEMANS

Science Research Associates, Inc.

LOUIS D. COHEN

University of Florida

HAROLD A. EDGERTON

Performance Research, Incorporated

MAX D. ENGELHART

Chicago City Junior Colleges

E. B. GREENE

Chrysler Corporation

J. P. GUILFORD

University of Southern California

JOHN A. HORNADAY

Houghton Mifflin Company

E. F. LINDQUIST

State University of Iowa

FREDERIC M. LORD

Educational Testing Service

ARDIE LUBIN

U. S. Naval Hospital, San Diego

SAMUEL MESSICK

Educational Testing Service

WILLIAM B. MICHAEL

*University of California,
Santa Barbara*

HOWARD G. MILLER

*North Carolina State, The University
of North Carolina at Raleigh*

M. W. RICHARDSON

Richardson, Bellows, Henry and Co.

P. J. RULON

Harvard University

C. L. SHARTLE

Ohio State University

KENDON SMITH

*The University of North Carolina
at Greensboro*

THELMA G. THURSTONE

*University of North Carolina at
Chapel Hill*

HERBERT A. TOOPS

Ohio State University

JOHN E. WILLIAMS

Wake Forest College

E. G. WILLIAMSON

University of Minnesota

DOROTHY ADKINS WOOD

*University of North Carolina at
Chapel Hill*

VOLUME TWENTY-FIVE, NUMBER ONE, SPRING, 1965

THE MINIMAL PHI COEFFICIENT AND THE MAXIMAL PHI

J. P. GUILFORD

University of Southern California

THE fact that inequality of means in correlated, dichotomized variables has a marked effect upon the size of phi, the product-moment coefficient of correlation, is well known. The greater the difference in means, the greater the restriction in the size of phi, in terms of possibility of achieving a value approaching 1.0. Some years ago, Ferguson (1941) pointed this out and developed an equation for determining the maximal value phi could achieve with specified marginal means, p_1 and p_2 .

Some investigators, wishing to make allowances for the effects of this kind of restriction, not wanting the relative sizes of phi to be determined by variations in means, as in using coefficients of correlation in factor analysis, have resorted to the simple, but questionable, procedure of using the ratio ϕ/ϕ_{max} as the index of relationship. The writer has seen no mathematical justification for this procedure. This procedure is even more questionable when the obtained phi is a negative quantity, for in such a situation, the restriction is in the opposite direction, i.e., a bias away from the other extreme of -1.0 .

It is the purpose of this note not only to point out the error in this operation but also to derive an equation for estimating the minimal value of phi, or the maximal negative value it may achieve, given a certain combination of means. It will be seen that the minimal phi is estimated by a different equation than that for estimating the maximal phi, although similar in form. A similar derivation will be given for the maximal-phi equation, differing from that given by Ferguson.

TABLE 1

Basic Contingency Table for the Computation of a Phi Coefficient

		Variable X		
		0	1	
Variable Y	1	β	α	p_i
	0	δ	γ	q_i
		q_i	p_i	

Table 1 shows the typical 2×2 contingency table, with cell proportions and marginal proportions. The marginal means and their complements are p_i and q_i for variable X, and p_i and q_i for variable Y. The standard formula for computing phi from data like these reads

$$\phi = \frac{\alpha\delta - \beta\gamma}{\sqrt{p_i q_i p_i q_i}} \quad (1)$$

For a given set of marginal proportions, phi is at a minimum when either α or δ is zero (or when both are zero). Let q_i be the smallest of the four marginal frequencies. Thus, q_i must be arbitrarily chosen, and it may be in any of the four positions. Let q_i be the marginal frequency corresponding to q_i , in the other variable. Thus, where q_i goes with a score of 0 in variable Y, q_i goes with a score of 0 in variable X. With q_i assigned to the score of 0 on Y, the cell frequency that needs to be zero to minimize phi must be δ . With δ equal to zero, the formula for phi reads

$$\phi_{min} = \frac{-\beta\gamma}{\sqrt{p_i q_i p_i q_i}} \quad (2)$$

Now with δ equal to zero, $\beta = q_i$ and $\gamma = q_i$. Substituting these quantities in (2), we have

$$\begin{aligned} \phi_{min} &= \frac{-q_i q_i}{\sqrt{p_i p_i q_i q_i}} \\ &= -\frac{\sqrt{q_i q_i}}{\sqrt{p_i p_i}} \end{aligned} \quad (3)$$

or

$$\phi_{min} = -\sqrt{\frac{q_i q_i}{p_i p_i}}$$

or

$$\phi_{min} = -\sqrt{\frac{q_i}{p_i}} \sqrt{\frac{q_i}{p_i}},$$

which is easier for computing purposes when the two quantities at the right are available in statistical tables (Guilford, 1956).

The maximal phi for a given set of marginal frequencies can be estimated by means of a similar equation, similarly derived. For this purpose, let p_i be the largest of the four marginal proportions, as Ferguson does. For phi to be at a maximum, either β or γ must be zero (or both). If p_i is the largest marginal proportion, q_i is the smallest. For the maximal phi, γ should equal zero. With γ equal to zero, $\alpha = p_i$ and $\delta = q_i$. With the product $\beta\gamma$ equal to zero and with p_i and q_i substituted in (1), we have

$$\begin{aligned}\phi_{max} &= \frac{p_i q_i}{\sqrt{p_i q_i p_i q_i}} \\ &= \frac{\sqrt{p_i q_i}}{\sqrt{p_i q_i}},\end{aligned}$$

which is identical with Ferguson's equation.

A comparison of the two formulas, for ϕ_{max} and ϕ_{min} is clearest when we use actual data. It will be seen that the two are not identical,

TABLE 2

Application of Equations for ϕ_{max} and ϕ_{min} to the Same Data, with Similar and Dissimilar Pairs of Corresponding Marginal Values

Case A
Similar Pairs

		Variable X		
		0	1	
Variable Y	1	.3	.5	.8 (p_i)
	0	.1	.1	.2 (q_i)
		.4 (q_i)	.6 (p_i)	

$$\phi_{\max} = \frac{(.6)(.2)}{(.4)(.8)} = .612$$

$$\phi_{\min} = -\frac{(.2)(.4)}{(.8)(.6)} = -.408$$

Case B
Dissimilar Pairs

		Variable X		
		0	1	
Variable Y	1	.5	.3	.8 (p_i)
	0	.1	.1	.2 (q_i)
		.6 (q_i)	.4 (p_i)	

$$\phi_{\max} = \frac{(.4)(.2)}{(.6)(.8)} = .408$$

$$\phi_{\min} = -\frac{(.2)(.6)}{(.8)(.4)} = -.612$$

and they cannot be identical except under the special circumstance in which $p_i = q_i = .5$, for then the two formulas are identical, except with opposite sign, $\sqrt{q_i/p_i}$ and $-\sqrt{q_i/p_i}$. Of course, when p_i also equals q_i , the outcomes are $+1$ and -1 . When both means are different from their complements, then it makes a difference whether p_i is greater than or less than q_i . With either equation, p_i is always greater than q_i (when they are unequal). Let us call the instance with p_i also greater than q_i case A, and the instance with p_i less than q_i , case B. Such cases are represented in Table 2.

In case A, p_i greater than q_i , ϕ_{max} is .612 and ϕ_{min} is $-.408$. In this case, the maximum phi can approach $+1$ more nearly than the

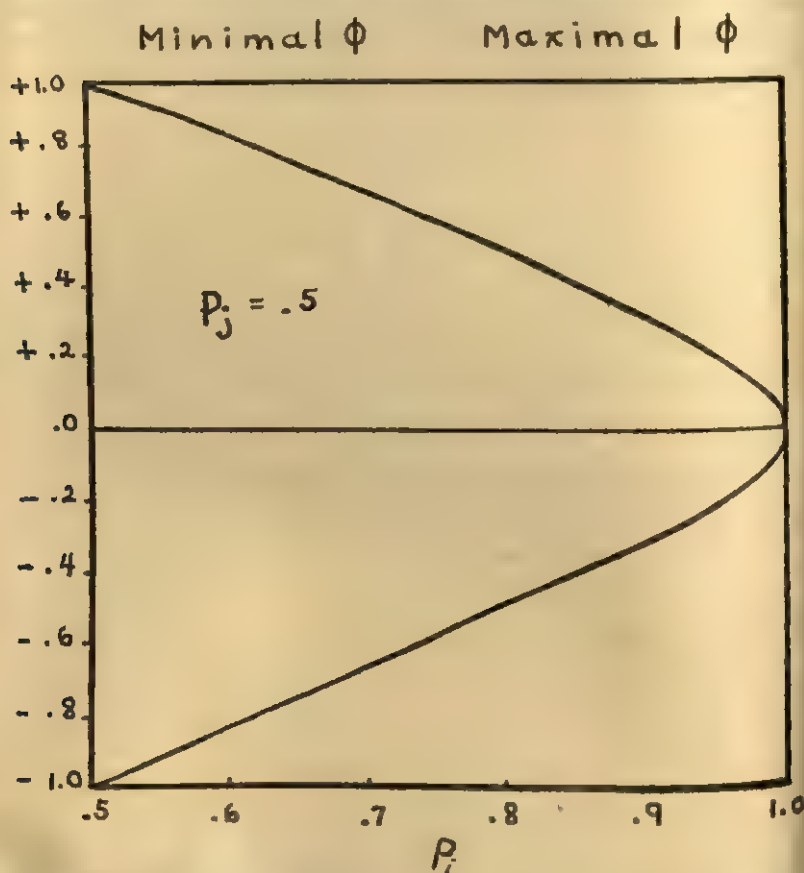


Figure 1. Maximal and minimal phi values as a function of p_i (the largest marginal proportion in a 2×2 contingency table) when p_i varies from .5 to 1.0 and when $p_i = q_i = .5$, where p_i is the proportion in the category corresponding to that for p_i , for the other variable.

Minimal ϕ Maximal ϕ

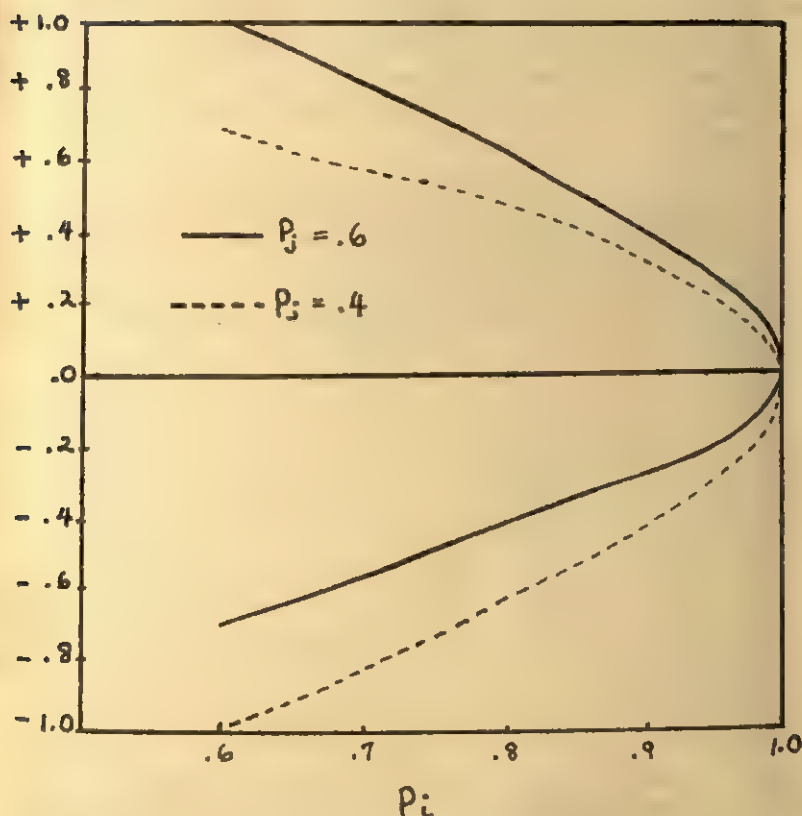


Figure 2. Maximal and minimal ϕ values as a function of p_i , when p_j is .6, also when p_j is .4.

minimum ϕ can approach -1 . In case B, with p_i less than q_i , the reverse is true, the quantities being $+.408$ and $-.612$, respectively. There is less restriction on a positive ϕ when the two pairs of corresponding marginal frequencies differ in the same direction. There is less restriction on a negative ϕ when the corresponding pairs differ in opposite directions.

The implications of these conclusions for the use of the ratio ϕ/ϕ_{max} are not clear, but it should be clear that with negative obtained ϕ 's, the ratio should be ϕ/ϕ_{min} , assigning a negative sign to the ratio. The general principle is that the range from ϕ_{min} to ϕ_{max} is usually asymmetrical about zero.

To show the more general picture, Figures 1 and 2 are presented.

In Figure 1, it is assumed that $p_i = q_i = .5$, and p_i is allowed to vary from .5 to 1.0. As indicated earlier, ϕ_{max} should equal ϕ_{min} under these conditions for all values of p_i . In Figure 2, however, two pairs of curves are shown; one pair for $p_i = .6$ ($q_i = .4$) and one pair for $p_i = .4$ ($q_i = .6$). The typical asymmetry when $p_i \neq q_i$ is illustrated by the two curves.

REFERENCES

- Ferguson, G. A. "The Factorial Interpretation of Test Difficulty." *Psychometrika*, VI (1941), 323-333.
- Guilford, J. P. *Fundamental Statistics in Psychology and Education* (3d ed.). New York: McGraw-Hill, 1956, Table G, Appendix B.

THE RELATIONSHIP OF PROBLEM-SOLVING STYLES TO THE FACTOR COMPOSITION OF TESTS¹

JOHN W. FRENCH

New College,
Sarasota, Florida

WHEN an ability or achievement test is constructed, its author usually assumes, or at least hopes, that his creation will measure the same thing for all examinees. An often talked about exception to this desirable quality of a test is the case of a fourth grader and a college student who take the same test of addition. We think such a test is one of reasoning for the younger subject, while it is one of speed for the older one. Some specific data on this appears in Balinski (1941). Large age differences, however, are not the only differences that affect the nature of the factor being measured by a test. Some tests of higher mental processes are solved in one way by some subjects and in another way by other subjects (Bloom and Broder, 1950, Ch. 3; Lucas, 1953). This means that the tests may be measuring different abilities for some subjects than for others, and so it follows that tests will have different correlations with one another for different kinds of groups of subjects. Therefore, since factor loadings of the tests depend on these correlations, "the factor loadings cannot be expected to be invariant from one population to a different population" (Thurstone, 1947, p. 360). It seems possible that the factorial composition of test problems involving higher mental processes often appears complex, not only because the problems require several different kinds of abilities in their solution, but also because they

¹ Grateful acknowledgment is made to Mr. Leighton A. Price, who skillfully and conscientiously conducted over 200 interviews, scheduled the testing, and helped in assembling the data used in this research.

Acknowledgement is also made to Educational Testing Service, where the study was carried out.

measure something different for examinees who solve them by using different methods. This study was initiated because an investigation of this situation seemed likely to lead to worthwhile information about individual differences in problem-solving styles, an appreciation of the different things a test measures for different subjects, and an understanding of the extent to which factor analysis reflects qualitative differences in individuals' reactions to a test as well as to differences in the nature of the tests themselves.

Briefly, the procedure for this study included the following steps. (1) The subjects took 15 tests which would later be examined by factor analysis. (2) In order to find out what problem-solving styles the subjects used, they were required to fill out a questionnaire about their background and their approach to test problems, and were interviewed while they solved items similar to ones they had faced on the tests. In all, more than 100 items of information about each subject were gathered. (3) As an aid in recognizing the important problem-solving styles in these data, the intercorrelations of the interview, the questionnaire, and the test variables were factor analyzed. (4) These variables, after being grouped and selected by the factor analysis, were used to divide the subjects into 17 different pairs of subsamples representing different problem-solving styles or different background characteristics that might be expected to affect the way in which a person solves problems. (5) All subsamples in the 17 pairs and also the whole group were then used in performing separate factor analyses of the same 15 tests. In all of these 35 factor analyses five factors were extracted and rotated. (6) Factor loadings and factor intercorrelations were compared for the two subsamples in each pair.

Selection of Tests

It was considered desirable to select tests known to have loadings on factors which are of general interest to researchers, which are involved in cognition rather than perception or pure speed, and, preferably, which have some demonstrated importance in intellectual achievement. A history of some confusion in the interpretation of a factor was regarded as an asset in this kind of study, since this left wide open the possibility that differences in problem-solving styles would affect factor loadings. It was also considered desirable to use some tests that straddle two or more of these factors. The five factors

that were selected and the five tests selected to represent them most purely are as follows.

<i>Factor</i>	<i>"Pure" test</i>
Verbal Comprehension	Vocabulary
General Reasoning	Mathematics Problems
Space	Cards
Induction	Letter Grouping
Visualization	Surface Development

A selection was made of ten other tests which would load these factors less purely or would straddle two or more of them. The 15 tests are named and described below. Following each test name, appear the code letters that will be used to refer to the test. Information about factor loadings previously found for all of the tests may be found in French (1951). A phrase indicating the factors usually found to be associated with each test is given at the end of the test description.

1. *Vocabulary, VOC*

A 5-choice synonym test of general vocabulary having a moderately wide range of item difficulties. 36 items, 7 minutes. Pure on Verbal Comprehension.

2. *Mathematics Problems, MAT*

A test of fairly conventional word problems all of which are solvable by arithmetic, but for a few of which the use of simple equations would help. 15 5-choice items, 15 minutes. Pure on General Reasoning.

3. *Cubes, CUB*

A Thurstone test in which each item presents 2 drawings of a cube with designs on its faces. Assuming no cube can have 2 faces alike, the subject indicates which items show 2 drawings that can be of the same cube and which ones show 2 drawings that cannot be of the same cube. 44 items, 5 minutes. Space with some Visualization and Induction.

4. *Punched Holes, PUN*

A Thurstone test in which each item includes successive drawings illustrating 2 or 3 folds made in a square sheet of paper. A drawing of the folded paper shows where 1 or 2 holes are punched in it. The subject draws the sheet as it would appear fully opened. 10 items, 5 minutes. Mainly Visualization.

5. *Letter Grouping, LET*

A form of Thurstone's test in which each item presents 4 sets of 4 letters each. The task is to find the rule which relates 3 of the sets to each other and to mark the one which does not fit the rule. 25 items, 3 minutes. Pure on Induction.

6. *Surface Development, SUR*

This is Thurstone's test in which there are drawings of solid geometrical forms. With each drawing there is a diagram showing how a flat pattern might be cut and folded so as to make the solid form. The subject is to indicate which lettered edges on the solid form correspond to numbered edges or folds in the pattern. 12 patterns totaling 60 items, 10 minutes. Mainly Visualization.

7. *Verbal Analogies, VER*

Each item consists of 2 words which have a certain relationship to each other followed by 5 numbered pairs of related words. The task is to circle the one of the numbered pairs which illustrates the same relationship as the original pair. 16 items, 10 minutes. Verbal Comprehension with some Induction and General Reasoning.

8. *Ship Destination, SHI*

A test by Christensen and Guilford in which the task is to use knowledge of the position of a ship in a diagram with respect to a port, knowledge of wind direction, ocean current, and direction of travel to compute effective distance to port following given rules. 48 items, 15 minutes. General Reasoning with some Space.

9. *Cards, CAR*

A Thurstone test in which each item gives a drawing of a card cut in some shape and sometimes having holes punched in it. To its right are 6 other drawings of the card, some merely of the card in different rotational positions and some of the card turned over. The subject indicates which ones are "like" the stimulus card (not turned over). 18 cards, 108 scorable units, 5 minutes. Pure on Space.

10. *Reading, REA*

Four paragraphs representing different kinds of reading were the basis on which the subjects responded to a total of 15 5-choice questions. 15 minutes. Verbal Comprehension with some General Reasoning.

11. *Concealed Figures, CON*

An adaptation by Thurstone of the *Gottschaldt Figures Test*. Each row of the test has a simple geometrical figure at the left side followed by 4 complex figures. The task is to mark each complex figure whether it does or does not contain the simple figure. 42 rows, 168 scorable units, 8 minutes. Space with several other factors.

12. *Figure Analogies, FIG*

One of 5 geometrical figures is to be selected as the fourth term in a proportion such that it bears the same relation to the third term as the second did to the first. 30 items, 5 minutes. General Reasoning, Induction, and Space.

13. *Spatial Orientation, SPA*

Each item presented 2 drawings of shore objects as seen over the prow of a boat. By comparing the position of shore objects with respect to the prow the subject was to judge the vertical, horizontal, or tilting motion of the boat between pictures. The subject reacted by choosing 1 of 5 responses, each of which showed a dot representing the first position of the prow and a dash representing the second. 54 items, 10 minutes. Two additional scorings of this test concerned separating the errors made in recognizing tilt from errors on other aspects of the test. Space and Visualization.

14. *Number Series, NUM*

Each item consisted of a series of seven 1- or 2-digit numbers. The eighth member of the series was to be selected from among the 5 choices offered. 30 items, 10 minutes. General Reasoning with some Induction.

15. *Marks, MAR*

A test by Thurstone in which each item gives 5 rows of places and gaps. In the first 4 rows one place is marked according to a rule. The task is to discover the rule and to mark one of the places in the fifth row accordingly. 20 items, 8 minutes. Induction with some Space and General Reasoning.

Administration of the Tests

The tests were administered to male students in the eleventh and twelfth grades of the college preparatory program at Princeton (New Jersey) High School and students enrolled in Princeton University. These subjects were paid for their participation in the experiment. A total of about 220 took the tests; 177 were used in the analysis.

The Interview and Questionnaire

Items like those in each of the 15 tests except Vocabulary and Ship Destination, were presented to each of the subjects individually within a few days after he took the tests. The subjects were asked to respond to each item, doing all their thinking aloud. In addition, the interviewer asked a few specific questions which seemed helpful in reaching a satisfactory understanding of the subject's methods or styles in solving each kind of problem presented to him.

The interviewing procedure was built up and continually revised during the processing of about 40 subjects whose test scores and interview protocols were not used in the analysis. During this period it was decided what kinds of things should be observed, what test items were best able to bring out a subject's characteristic behavior, and what questions should be asked in order to facilitate interpretation of the subject's methods for solving the problems that were set for him. Although the decision on the kinds of behavior to be recorded was decided entirely on the basis of what seemed especially distinctive during the interviews, there was, nevertheless, a relationship between the kinds of behavior observed in this study and characteristics of problem-solving behavior reported in the literature and discussed in the next section.

Items from the Vocabulary and Ship Destination tests were not used in the interview, because no satisfactory differences in methods of working on these tests could be observed.

During the course of the development of the procedure, some 20 interviews were carried out jointly by the writer and the interviewer, who made independent judgments of the subjects' responses. When substantial disagreement occurred, interview test items were changed or eliminated, the nature of the ratings were redefined, or special questions for the subject were incorporated in the interview procedure.

The interview yielded 105 ratings or variables. Some of these

were necessarily dichotomous because of their essentially qualitative nature; some of them were ratings on a 3-point to 7-point scale or were enumerations of specific ideas or words spoken by the subject. A multi-point scale was used for a rating whenever possible in order to permit conversion to a dichotomy with which the group of subjects could be split into two nearly equal divisions.

After the interview, the subjects completed a questionnaire asking them about their school and home interests, academic interests and background, their abilities, their ambitions, and the way they themselves thought that they solved problems. While most of these questions did not pertain directly to problem-solving styles, they concerned characteristics of the individuals that might be expected to relate to problem-solving styles. The questionnaire yielded an additional 32 variables. Most of these consisted of an item that could be checked or not checked or an activity that could be named or not named. That is, most of the variables yielded by this questionnaire were dichotomies, many of which, however, split the total group of subjects into poorly balanced subgroups.

The Selection of the Experimental Dichotomies

Experimental dichotomies constructed from single variables or combinations of variables were to be used for dividing the total sample of subjects into pairs of contrasting subsamples. Since the interview and questionnaire provided a total of 137 variables, it seemed neither wise nor possible to decide about the dichotomies on the basis of simple inspection or common sense.

All variables which were not already dichotomies were dichotomized as closely as possible to the center of their distributions. The 15 test scores were also dichotomized. For the resulting 152 dichotomies the matrix of tetrachoric intercorrelations based on 177 cases was computed. Variables shown by these correlations to be mere repetitions of others and also some variables for which the split was particularly unbalanced were now dropped from further analysis. The remaining 125 variables were factored. While the ratio of 177 cases to 125 variables is extremely small, factor analysis was considered to be a reasonable procedure in this instance because it was being employed merely as a technique to suggest groups of variables that could be combined into short, rough scales for separating the subjects into pairs of subsamples. Twenty-five principal component

factors were extracted using communalities estimated by iteration. The decision to stop extracting factors was based on the size of the eigenvalues. The factors were then rotated orthogonally by the varimax method.² While the total group of subjects could have been divided into subsamples on the basis of any one of the variables, each of these 25 factors was thought to represent a particularly fruitful independent basis for the division. Actually only the 17 factors which seemed to have psychological meaning were used for this purpose.

This article presents the analysis of subsample Pair Numbers 2, 3, 8, and 15. The heading of each analysis table (Tables 3 through 6) contains the name of the factor (pair). A note at the bottom of each table gives the names of the 3 or 5 variables forming the scale used to divide the total group of subjects into two subsamples. These variables are ordinarily the ones with highest loadings on the one of the 17 factors from which the pair was derived, although variables highly correlated with one another were sometimes avoided.

A consideration of the names of the 17 factors given in a later section will show that many of the subsample divisions, while being developed from *orthogonal* factors, nevertheless seem to fall loosely into a category that can be called "systematizing" or "analyzing vs. scanning." All four of the pairs to be discussed at length here are included in this category.

The prevalence of problem-solving styles that seem concerned with systematizing vs. scanning suggests that there may be a relationship between the problem-solving styles observed in this experiment, particularly for the spatial tests, and the cognitive style of focusing vs. scanning which has been found to effect perception of size and susceptibility to illusions (Gardner et al., 1959; Gardner, 1961; Gardner and Long, 1962). For some of the pairs the observed problem-solving style can best be termed "analyzing" in contrast to a global way of perceiving. This sounds like a dimension described by Witkin et al. (1962). It has usually been found to be positively correlated with intellectual ability (Witkin et al., 1954, 1962; Gardner, Jackson, and Messick, 1960). The means of the subsamples on the verbal and reasoning tests, VOC, NUM, MAR, LET, and MAT, given at the bottoms of the tables consistently confirm

² All computations were programmed and carried out by Dr. David R. Saunders.

this positive relationship between analyzing and "intelligence." However, the analytic approach was not found to be consistently successful for spatial tests or even for the particular tests with which the analytic behavior was observed. In particular, the subsamples using analytical behavior on concealed Figures (Pair 8) and Cubes (Pair 15) made lower mean scores on those tests than did the subsamples using global behavior. Since Concealed Figures is very similar to the test that Witkin and his associates used for measuring Analytic Attitude or Field-Independence, there appears some doubt that the behavior being observed here is, in fact, the same as that described by Witkin.

Factorization of the Subsamples

After the 17 pairs of subsamples had been defined, the matrix of intercorrelations of the 15 test scores for each subsample and for the total sample were factored. In all cases five factors were extracted. To stipulate the same number of factors in advance for all 35 analyses effected a considerable saving in cost. The number five was used because the tests had been chosen so as to include variance from five known factors. Examination of the eigenvalues and residuals gave no indication that a different number of factors would have been preferable.

The first rotation for each of the 35 factorizations was by varimax to an orthogonal solution. Since precise comparisons of factor loadings were of central importance to this study, a rotational technique which insured congruency of the factors was essential. To achieve this, four of the factors were rotated to specified marker variables defining the positions of the factors. With one exception, the marker variables selected were those with the highest loadings on the four psychologically meaningful factors in the varimax solution for the whole group. The Concealed Figures test, CON, actually had the third highest loading on a factor recognized as a Space-Visualization factor, but this test was not used as a marker, because it has not always been clearly identified with Space or Visualization.

A quartimax rotation was applied first to produce the best orthogonal rotation to the four sets of marker tests. Factor E was then permanently set orthogonal to these four. A patterned oblimax rotation was then applied to Factors A, B, C, and D to bring these factors as close as possible to the marker tests. With the exception

TABLE 1

Factor Loadings for the Total Sample after Rotation

Test Name	Code	Factor Loadings				
		A	B	C	D	E
1. Vocabulary	VOC	-.01	.68*	-.00	-.11	.02
2. Mathematics Problems	MAT	.11	.22	.41*	-.08	-.15
3. Cubes	CUB	.46	.17	-.14	.01	.21
4. Punched Holes	PUN	.42*	-.01	.01	.15	-.07
5. Letter Grouping	LET	.10	.07	.25	.08	.45
6. Surface Development	SUR	.80*	.08	-.10	-.17	-.06
7. Verbal Analogies	VER	.03	.60*	-.03	-.02	-.04
8. Ship Destination	SHI	.06	.16	.05	.31*	-.02
9. Cards	CAR	.57*	-.08	.03	-.01	.10
10. Reading	REA	.02	.62*	-.08	.04	.03
11. Concealed Figures	CON	.43	.02	-.02	.15	.11
12. Figure Analogies	FIG	.40	.06	-.02	.15	.06
13. Spatial Orientation	SPA	.37	.17	-.09	.12	-.05
14. Number Series	NUM	-.10	-.11	.63*	.02	.04
15. Marks	MAR	-.12	-.08	-.00	.52*	.00

* Marker tests used to determine the position of rotation.

of Factor E, the resulting rotations were oblique, often extremely oblique, correlations between factor axes ranging from .28 to .80.

For the total sample, Table 1 shows the loadings for the oblique rotation and Table 2 shows the intercorrelations of the factor axes. The factors have been named as follows.

Factor A, (marked by PUN, SUR, and CAR). Space-Visualization. In selecting the 15 tests for this study, it was intended that Space and Visualization would appear as separate factors, but they did not do so. As could be seen from factor plots made from the data in Table 1 and from the data for some of the subsamples, the spatial and visualization tests are intermingled. Therefore, failure to isolate the two factors is not merely a defect of the rotation.

TABLE 2

Correlation of Factor Axes for the Total Sample

Factors	A	B	C	D
B	.44			
C	.54	.69		
D	.67	.60	.65	
E	.00	.00	.00	.00

Factor B, (marked by VOC, VER, and REA). Verbal Comprehension.

Factor C, (marked by MAT and NUM). Mathematics. MAT was intended to associate with SHI to define the General Reasoning factor as it has in many studies carried out in Guilford's laboratory (Green et al., 1953 etc.) but it did not do so. Its association with NUM is reasonable, but the factor seems to be one of number facility or mathematical content rather than one of pure reasoning.

Factor D, (marked by SHI and MAR). Reasoning. If MAR and LET had appeared on the same factor, it could have been interpreted as induction. However, since SHI is a reasoning test but probably not a test of inductive reasoning, it will serve the purposes of this study to call Factor D a reasoning factor without specifying what kind of reasoning is involved.

Factor E, (orthogonal to A, B, C, and D). No interpretation. LET has the highest loading, but data for making an interpretation are very slight indeed.

The matrix of factor loadings in Table 1 may be compared with the results given here for four of the pairs of subsamples. Only selected figures are given for the subsamples; otherwise the mass of numbers would make comparisons very cumbersome. The table for each pair (Tables 3, 4, 5, and 6) presents the following data for each subsample.

1. Immediately under the letters *A*, *B*, *C*, *D*, and *E* are given those loadings that are .20 or higher on each of the five factors.
2. Given also are *all* loadings for the test involved in defining the subsamples. These appear in parentheses when less than .20.
3. Next come the intercorrelations of factor axes.
4. At the bottom of the tables is given the subsample mean and standard deviation for some of the tests including those involved in defining the subsamples. These are given to provide a fuller understanding of the kind of subjects falling into each subsample.

Discussion of the Subsample Results

To save space data are given here for only four of the 17 pairs of subsamples that were analyzed. However, following a discussion of the findings for Pair Numbers 2, 3, 8, and 15, the other 13 pairs are named and a brief note on the findings for each of them is given.

TABLE 3
Results for Pair 2—Uses Rule in Solving Cards Items

Does not use rule, $N = 109$										Uses rule, $N = 68$									
A					B					C					D				
E					F					G					H				
I					J					K					L				
M					N					O					P				
Q					R					S					T				
U					V					W					X				
Y					Z					AA					AB				
AC					AD					AE					AF				
AG					AH					AI					AJ				
AK					AL					AM					AN				
AO					AP					AQ					AR				
AS					AT					AU					AV				
AW					AX					AY					AZ				
BA					BB					BC					BD				
BE					BF					BG					BH				
BG					BI					BJ					BK				
BL					BM					BN					BO				
BN					BP					BQ					BR				
BS					BT					BU					BV				
BU					BW					BX					BY				
BW					BX					BY					BZ				
BY					BZ					CA					CB				
CA					CB					CC					CD				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				
CK					CM					CN					CO				
CM					CN					CO					CP				
CN					CO					CP					CQ				
CO					CP					CQ					CR				
CP					CQ					CR					CS				
CQ					CR					CS					CT				
CR					CS					CT					CU				
CS					CT					CU					CV				
CT					CU					CV					CW				
CU					CV					CW					CX				
CV					CW					CX					CY				
CW					CX					CY					CZ				
CX					CY					CZ					CA				
CY					CZ					CA					CB				
CZ					CA					CB					CC				
CA					CB					CC					CD				
CB					CC					CD					CE				
CC					CD					CE					CF				
CD					CE					CF					CG				
CE					CF					CG					CH				
CF					CH					CI					CJ				
CH					CJ					CK					CL				
CJ					CK					CM					CN				

Pair 2, Uses Rule in Solving Cards Items. Division of the sample into a pair of subsamples according to whether or not the subjects used a rule for analyzing Cards items produced no appreciable differences in the factor loadings or any large differences in the intercorrelations among factors. The analysis of this pair is included simply as an example of one of six Pairs (Nos. 2, 4, 5, 6, 11, and 20) showing substantially negative results except for certain general effects noted below. The reason for a negative result for this pair of subsamples, or for any other pair of subsamples, cannot, of course, be fully determined. The implication here is that the Cards test measures substantially the same thing whether or not the subject uses a rule in responding to the items. It is possible, however, that the interview simply failed to make any reliable discrimination with regard to behavior on this particular test.

Pair 3, Emphasizes Relationships in Number Series Items. This division, based on the use of relationships in solving Number Series problems, NUM, has affected Factor C, the mathematics factor. In going from the group not using relationships to the group using relationships, the loading of the Mathematics Problems test, MAT, declines sharply on Factor C. Since the division of subsamples is based on NUM, a reasonable interpretation of this occurrence is that NUM moved away from the mathematics factor and is now forming a new Factor C, which appears inductive in nature. This leaves no factor that can be interpreted as Mathematics. This is one of several pairs of subsamples showing that a systematic approach to a test tends to move the test away from the factor or other tests with which it is usually associated.

Pair 8, Uses Details of Figures in Concealed Figures Items. The pattern of factor loadings is essentially unchanged by this division. However, going from not using details to using details, the intercorrelations of factors show that Space-Visualization has moved away from (that is, has become less correlated with) the others; in particular it has moved away from Verbal Comprehension. This is one instance of a very consistent occurrence found in many of the pairs. In pairs of subsamples where some systematic approach was used for a test (Pairs 2, 3, 5, 6, 7, 8, 9, 11, 15, 16, and 17) the correlation between Space-Visualization and Verbal Comprehension *always* goes down, sometimes drastically. In such cases the correlation between Space-Visualization and Mathematics also *always* goes down, ex-

TABLE 4
Results for Pair 3—Emphasizes Relationships in Number Series Items

Many relationships, $N = 83$									
A	B	C	D	E	A	B	C	D	E
SUR .71	REA .71	NUM .58	MAR .49	LET .40	SUR .84	VOC .76	NUM .63	MAR .56	LET .38
CAR .51	VOC .64	MAT .48	CON .26	PUN - .24	CAR .72	VER .56	LET .34	SHI .34	CON .36
PUN .38	VER .60	SPA - .24	SPA .24	(NUM .04)	CON .55	REA .56	MAT .29	(NUM -.03)	CUB .24
CUB .37	SPA .32	CUB - .23	SHI .23		CUB .55	MAT .25			SPA -.21
FIG .32	(NUM -.13)	LET .21	LET .23		PUN .52	CUB .20			(NUM .01)
CON .32			(NUM .09)		FIG .49	(NUM -.10)			
SPA .23					SPA .49				
(NUM -.15)					LET .21				
					(NUM -.06)				

Few relationships, $N = 94$									
A	B	C	D	E	A	B	C	D	E
SUR .71	REA .71	NUM .58	MAR .49	LET .40	SUR .84	VOC .76	NUM .63	MAR .56	LET .38
CAR .51	VOC .64	MAT .48	CON .26	PUN - .24	CAR .72	VER .56	LET .34	SHI .34	CON .36
PUN .38	VER .60	SPA - .24	SPA .24	(NUM .04)	CON .55	REA .56	MAT .29	(NUM -.03)	CUB .24
CUB .37	SPA .32	CUB - .23	SHI .23		CUB .55	MAT .25			SPA -.21
FIG .32	(NUM -.13)	LET .21	LET .23		PUN .52	CUB .20			(NUM .01)
CON .32			(NUM .09)		FIG .49	(NUM -.10)			
SPA .23					SPA .49				
(NUM -.15)					LET .21				
					(NUM -.06)				

Note: Pair 3 was defined as subjects who were marked plus on 3 or more of the following vs. all others:

86. Starts number series items by observing relationships;

88. Tries high number of relationships;

90. Does not try patterns that skip or do not account for all members;

93. Tries multiplying and dividing rather than only adding and subtracting;
94. When asked, reports trying relationships before scanning.

cept for Pairs 3 and 15 where there was no mathematics factor in one of the subsamples. This finding suggests that the subjects who attempt to solve spatial or other problems by ordinary common sense, perhaps by simple inspection until a solution seems to offer itself, succeed pretty much according to their general intelligence as measured by verbal and mathematical tests. For those subjects, therefore, the correlation between Space-Visualization and the verbal or mathematics factors is high. On the other hand, the subjects who use a system for solving problems have succeeded in developing some specialization of their abilities, and so the correlations drop down.

Pair 15, Analytic Approach to Cubes Items. This division based on the Cubes test produced rather dramatic results. The loading of Cubes dropped from .52 to .07 on Space-Visualization. Apparently the kind of analytic behavior toward Cubes, which was observed in the interview, completely destroys the capacity of that test to measure spatial ability. At the same time, the tests loading Factor B show that Verbal Comprehension and Mathematics have collapsed into one factor. The new Factor C, loading Letter Groups and Number Series, seems inductive in nature. It is this new factor that now carries the principal loading of Cubes, suggesting that the analytic behavior of the subjects has changed Cubes from a test of space to a test of induction. This finding seems to confirm the results of an earlier study (French, 1957), where most of the conventional factors were clear, but where space and induction tests fell on one dimension. This happened with West Point Cadets, who may have approached the tests with a particularly analytical attitude.

The verbal-mathematical composite in Pair 15 was a surprise. It was not forced by the limitation on the number of factors extracted, because Factor E has so small a variance that it could have provided room for a fifth well-defined factor. The collapse of the mathematical and verbal factors is probably related to an occurrence which is consistent in all but one of the pairs where a systematic approach is used for the solution of problems (unless there is no mathematics factor as occurs for Pairs 3 and 15). These Pairs are Nos. 2, 5, 6, 7, 8, 9, 11, 16, and 17. In all but one of them, Pair No. 8, Mathematics and Verbal Comprehension are more highly correlated for the systematic subsample than they are for the non-systematic subsample. For Pair 15, Mathematics and Verbal Comprehension

TABLE 5
Results for Pair 8—Uses Details of Figures in Concealed Figures Items

Does not use details, N = 81										Uses details, N = 96														
A					B					C					D					E				
SUR	.73	VER	.68	NUM	.65	MAR	.64	MAT	.31	SUR	.80	VOC	.71	MAT	.54	MAR	.38	LET	.51					
CAR	.45	VOC	.63	MAT	.36	SHI	.31	LET	-.31	CAR	.66	REA	.64	NUM	.48	SHI	.32	CUB	.30					
CON	.40	REA	.58	LET	.35	(CON	.06)	PUN	.21	CUB	.58	VER	.55	LET	.21	FIG	.31	FIG	.28					
FIG	.39	MAT	.28	CUB	-.23	(CON	.06)	(CON	.06)	PUN	.46	(CON	.03)	(CON	-.06)	CUB	-.20	(CON	.17)					
PUN	.39	(CON	.02)	(CON	.04)					CON	.45					(CON	.19)							
SPA	.38									FIG	.34													
SPA	.37									SPA	.34													
										LET	.23													
					A					B					C									
					.59					.66					.69									
					.73					.64					.64									
					M.					S.D.					S.D.									
					30.7					14.3					11.6									
					21.6					8.5					7.9									
					24.3					4.5					4.4									
					9.2					4.1					3.5									
					14.5					3.7					3.9									
					12.3					3.3					2.9									
					72.1					22.2					19.2									

have simply become correlated in the extreme; they have collapsed into one factor. Thus, for the systematic person, Space-Visualization is relatively independent, while the verbal and mathematics factors are relatively close together. As discussed in connection with Pair 8, the effect on Space-Visualization can be regarded as a development of specialized techniques or approaches to spatial problems. The systematic person is also likely to attempt the solution of verbal and mathematical problems by the development of specialized techniques. Since verbal and mathematics problems are both symbolic in nature, the closing together of these factors can, perhaps, be understood by supposing that the systematic subjects can develop symbolic approaches or attitudes which help them jointly in solving both verbal and mathematical problems. Such techniques, being common to the symbolic verbal and mathematics factors, draw these two factors together, while they may be quite different from the techniques developed for the visible, concrete, non-symbolic spatial problems, thus leaving the spatial factor relatively isolated in the factorial space.

Brief Notes on Other Pairs

Pair 1, Finds Less vs. More Trouble with Mathematics Problems. Because the Mathematics Problems test was too easy for the mathematically able subsample, it failed to discriminate and no mathematics factor appeared.

Pair 4, Likes Verbal Subjects vs. Likes Science. No important difference.

Pair 5, Uses Hypotheses for Letter Groups Items. No important difference.

Pair 6, Uses Hypotheses for Marks Items. There was a drop in correlation between the reasoning factor, which contains Marks, and Verbal Comprehension.

Pair 7, Uses Reasoning for Spatial Orientation Items. Spatial Orientation dropped from the Space-Visualization factor and became loaded on the Reasoning factor. The Reasoning factor's correlation with Verbal Comprehension declined sharply.

Pair 9, Considers Each Reading Item Alternative. The loading of Reading on the Verbal Comprehension factor was reduced.

Pair 11, Studies Relationships in Verbal Analogies Items. With-

out apparent reason there was a sharp decline in the loading of Punched Holes on Space-Visualization.

Pair 12, Notes Points while Reading in the Reading Test. The loading of Reading on the Verbal Comprehension factor was reduced. Neither of the behaviors indicated in Pairs 9 or 12 has much relationship to the mean score on the Reading test.

Pair 13, Art or Music Activities. Without apparent reason Ship Destination has a low loading on the Reasoning factor for the non-art subsample.

Pair 14, Few Visualization Indications. No important difference except for a lowering of correlations between the Reasoning factor and others.

Pair 16, Geometrical Approach to Punched Holes and Surface Development. Sharp but not consistent changes in the loadings on Space-Visualization were observed.

Pair 17, Can List Many Hypotheses for Letter Grouping and Marks Items. Unlike Pair 6 the Reasoning factor and Verbal Comprehension became more highly correlated.

Pair 20, Faster Rejection than Acceptance of Concealed Figures Items. No important differences.

Factors 10, 18, 19, and 21-25 were not used in creating pairs of subsamples.

Summary of Findings and Conclusions

There seems to be extensive evidence here that tests, even simple "pure-factor" tests, do not measure the same things for all people. The kind of behavior most readily observed in someone taking these tests was the use of some kind of reasoned or systematic approach as contrasted to less orderly scanning and visualizing, with reliance on common sense. This overall difference in problem-solving styles is the one emphasized by Bloom and Broder (1950, see especially Appendix), and may be related to Gardner's (1959) Focusing vs. Scanning or to Witkin's (1962) Analytic Attitude or Field-Independence. For some of the tests, differences of this kind in the test-taking behavior have no relation to the test's factorial content. For a few tests, the principal loading was strengthened. Most often, however, the use of a system in solving a test reduced the usual factor loading of that test. This happened more regularly for spatial or visualization tests than for reasoning or verbal tests. These find-

ings are understandable in terms of the view that tests are efforts on the part of the test constructor to set up standard tasks for the subject. In some cases the systematic approach to a test may eliminate some random behavior, thereby increasing reliability and increasing the expected test loadings. More often, however, the system enables the student to respond to the test items after undergoing a task partly or entirely different from the one that the test constructor had in mind. In such cases the expected factor loadings of the test decline or vanish altogether.

Some rather general findings were also observed with regard to the intercorrelations of the factors. As compared to the factor correlations for subjects who did not use systematic problem-solving methods, the correlations for the more systematic subjects showed the Space-Visualization factor to be further off by itself and Verbal Comprehension and Mathematics to be closer together. An explanation for this was attempted by suggesting that the more reasoned problem-solving style permitted the discovery through a reasoning process of specialized techniques of a spatial nature, while the same reasoning process directed the subjects toward a common means for solving symbolic problems with either verbal or mathematical content.

Another way of thinking about changes in what tests measure is to consider as "moderator variables" the dichotomies that were used to divide the total sample into subsamples. Moderator variables are variables whose values are associated with different amounts of correlation between two other variables; thus they seem to moderate or regulate the correlation between the other variables (Saunders, 1956). Systematizing is a variable that was found to be associated with lower correlations between the test being systematized and other tests ordinarily loading the same factor. This caused a decrease in the loadings of the test in question on that factor. Going beyond the test itself, the finding was that systematizing on any test was associated with lower correlations between spatial and other kinds of tests and higher correlations between verbal and mathematical tests. These changes in correlations resulted in corresponding changes in correlations among the factors. The above-mentioned speculations about the specialization of spatial ability and the use of symbolic reasoning to solve both verbal and mathematical problems now apply as explanations of the action of systematizing as a mod-

erator variable. The logical sequence is this: systematizing is a tendency which leads a person to use specialized or symbolic thought processes; this changes what the tests measure and, consequently, affects the correlations between the tests.

REFERENCES

- Balinski, B. "An Analysis of the Mental Factors of Various Age Groups from Nine to Sixty." *Genetic Psychology Monographs*, XXIII (1941), 191-234.
- Bloom, B. S. and Broder, L. J. *Problem-solving Processes of College Students*. Chicago: University of Chicago Press, 1950.
- French, J. W. "The Description of Aptitude and Achievement Tests in Terms of Rotated Factors." *Psychometric Monographs*, (1951), No. 5.
- French, J. W. "The Factorial Invariance of Pure-factor Tests." *Journal of Educational Psychology*, XLVI (1957), 93-109.
- Gardner, R. W. "Cognitive Controls of Attention Deployment as Determinants of Visual Illusions." *Journal of Abnormal and Social Psychology*, LXII (1961), 120-127.
- Gardner, R. W., Holzman, P. S., Klein, G. S., Linton, Harriet, and Spence, D. R. "Cognitive Control." *Psychological Issues*, I (1959), No. 4.
- Gardner, R. R., Jackson, D. N., and Messick, S. "Personality Organization in Cognitive Controls and Intellectual Abilities." *Psychological Issues*, II (1960), No. 8.
- Gardner, R. W. and Long, R. I. "Control, Defense and Centration Effect: A Study of Scanning Behaviour." *British Journal of Psychology*, LIII (1962), 129-140.
- Green, R. F., Guilford, J. P., Christensen, P. R., and Comrey, A. L. "A Factor-Analytic Study of Reasoning Abilities." *Psychometrika*, XVIII (1953), 135-160.
- Lucas, C. M. "Analysis of the Relative Movement Test by a Method of Individual Interviews." Princeton, N. J.: Educational Testing Service, 1953.
- Saunders, D. R. "Moderator Variables in Prediction." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVI (1956), 209-222.
- Thurstone, L. L. *Multiple-factor Analysis*. Chicago: University of Chicago Press, 1947.
- Witkin, H. A., Dyk, R. B., Faterson, Hanna F., Goodenough, D. R. and Karp, S. A. *Psychological Differentiation*. New York: Wiley, 1962.
- Witkin, H. A., Lewis, Helen B., Hertzman, M., Machover, Karen, Meissner, Pearl B., and Wapner, S. *Personality through Perception*. New York: Harpers, 1954.

INTRIGUING PROBLEMS OF DESIGN IN PREDICTING COLLEGE SUCCESS*

DANIEL D. FEDER

San Francisco State College

For many years efforts to predict success in college have been focused on devising a combination of prior academic records, aptitude and achievement tests, and certain subfactors in the environment which would yield the highest possible coefficient of correlation with a single criterion presumed to be the measure of college success—the grade point average. Such attention has centered on the prediction of freshman year achievement, since much of the motivation for such studies came from institutions of higher education concerned with selective admissions. Another major motivation was the effort to obtain guidelines for counselors to use in appraising the high school student's chances for success in the college of his choice. For all institutions of higher education there has certainly been a common motivation which is the identification of student potential and the guidance of the student to an achievement level that is appropriate with that potential.

Narrowness of Objective in Predictive Validity Studies

The fact remains that the rather vast literature on this subject has been directed toward a very narrow objective when one considers that the sole criterion to be predicted is grade point achievement or even the broader one of college graduation. Out of the very considerable literature on the subject, some combination of prior

* A paper presented to the Symposium entitled "Measurement and Prediction in the College Admissions Process," Fifteenth Annual Conference on Educational Research sponsored by the California Advisory Council on Educational Research (CACER) in Los Angeles, November 15, 1963.

marks such as the high school grade average has consistently been the best single predictor. Combination of these marks with an aptitude test supplying a quantitative and a verbal score, with a test of reading comprehension, and with a general achievement test has yielded coefficients in the high .60's. This result still leaves a large area of unpredicted variability.

Some researchers have expressed considerable frustration as they have striven to increase these coefficients of correlation. Yet, realistically, the obtained results are quite significant if one but bears in mind that the inherent unreliability of the criterion (grade point averages as computed from teachers' marks) is such as to set the encountered limits on the prediction correlations no matter how highly reliable some of the prediction instruments have become.

Research concerning Changes in Students during Educational Experiences

A growing body of knowledge in recent years deals with changes which take place in students as a result of their exposure to the educational experience. The survey by Jacob (1957) suggests that four years of higher education have little effect upon the values of college students. The Eddy (1959) study which concerned itself with the impact of a particular educational and social environment upon the character of students likewise produced a rather dismal picture. A series of studies emanating from the University of California's Center for Higher Education (Heist and Webster, 1960) suggests the possibility that students' achievements may bear direct relationship to the kind of intellectual and social environment which they encounter on a campus and their own personal perceptions of the congruence of such environment with their own desires. Such relationships have been found to be a factor in mortality rates and perseverance. Iffert's (1958) study of some 8000 college students in a national sample was concerned primarily with withdrawals and transfers. Such factors as financial difficulties, lack of interest in studies, requirements of military service, marriage, etc., which may have influenced a severance action certainly can be seen as having had impact also upon achievement in the college setting. Yet, relatively little attention has been given to the influence of such factors in predicting success in college as measured by grade point achievement.

Lest it be assumed that undue emphasis is being placed upon the importance of the grade point average as a major or even the sole criterion of success in college, let us at least admit the fact that it still represents one of the most readily tangible and quantifiable factors and hence becomes the center of our focus. In point of fact, the efforts to sharpen predictive instruments and subsequently to make specific predictions of achievement in courses are tantamount to putting a razor edge on a hoe. The ultimate uses to which the findings will be put do not warrant the kind of exactitude which is being sought. With the knowledge already available on prediction it is likely that for the uses of counseling, motivation, and even the selection of students for admission to selective higher institutions, adequate results could be obtained by dividing test scores into quartiles and using similar categories for college achievement. The resultant experience tables or contingency coefficients will yield results sufficiently meaningful for the purposes for which they are to be used.

An Examination of the Complexity of the Criterion Variable

Some of the more recent work in prediction, such as that of Holland and the University of California studies cited before, indicate the complexity of the achievement variable and focus sharply on the oversimplification of it in much of the research during the past 25 years. There is ample evidence that college grades are affected by noncognitive factors which are major operants in the college student's gestalt, but little or no attention of researchers has been directed toward them. Examination of a few of these factors may suggest needed directions and designs for future studies in predicting college success.

Level of aspiration has long been recognized as a central motivating force in behavior. Yet account of the level of aspiration has not been taken in attempting to describe some of the unpredicted variance of the low ability—high achiever. Holland's (1963) findings in this regard suggest that a particular college atmosphere, viz., the competitive versus the non-competitive campus, has impact on this operation. In the same vein many questions may be posed: what are the student's aspirations with reference to what he seeks from the college experience in addition to the formal aspects of education? What degree of self actualization (although he may not call it that)

is important to the student? What opportunities for social life and social development does he see as germane to his educational experience? What is the relative importance he attaches to grades themselves as compared with the opportunities for intellectual exploration, challenge, and excitement? Is he willing to challenge his teachers even though he knows that the campus tradition discourages such challenge in favor of a more conventional grade-getting activity? Parallel to these considerations is the fact that although the prediction instruments (cognitive measures) may appraise with accuracy and with validity the intellectual power of the individual, they may take no account at all of the adjustive and emancipative challenges with which the student is grappling as a characteristic developmental task in the particular point of time and experience which is associated with the college-going years.

Criterion Research in Relation to the Self Concept

Recognizing the limitations of the grade point average as a measure of the full impact of the college experience, we have been exploring for a number of years the measurement of the self concept and some of the ways in which it changes among college students. I shall cite here only a few of the observations from a series of studies (Ross, 1955; McKee, 1958; Miller, 1960) in which a variety of instruments for measuring the self concept changes were used. It is our belief that the evidence from these studies suggests that changes in the self concept may represent one of several valid noncognitive factors which are important aspects of college achievement broadly defined. In this broad definition, earned grades are only one part of the total achievement complex. But because they are a part of a dynamic gestalt they are, in turn, influenced by these noncognitive factors which in a very real sense are both cause and effect in the college prediction pattern.

We have observed that both low and high ability students who are relatively low achievers (although passing and earning grade point averages which would ultimately result in graduation) showed a higher altruistic self concept orientation than did high achieving students. That is to say, they were more concerned with their relationship as persons to a larger social group or to even a larger social order. Since they were more susceptible to peer groups norms, they may have been influenced by the "gentlemen's C" concept. Higher egocentrism was correlated with higher achievement. Moreover, us-

ing conventional terminology, one may say that egocentrism was characteristic of the over achiever. What degree of variance can be accounted for by this factor has yet to be determined.

In a study which followed a sample of college students through a four year span, altruistic orientations of the self concept were enhanced during the four years. A group who dropped out of formal higher education during the freshman year, although successful by conventional academic standards, retested four years later also, had regressed toward even greater egocentrism than manifested when they were tested as freshmen. Although these findings are in marked contrast to the observations of Jacob, they further suggest implications for prediction which have yet to be probed.

With the knowledge that college grades are susceptible to self concept orientation, it is also pertinent to point out that specifics in the college experience may influence self concept changes. When an instrument which yielded both a real and ideal self concept measurement, with a correlated discrepancy score was used, it was observed that certain types of curricular patterns may operate to reduce the discrepancy score. For example, exposure to a Science-Humanities emphasis curriculum was more influential than exposure to a Science-Social Science emphasis. Similarly, students in a Fine Arts major emphasis experienced even sharper reduction in the ideal self concept. In this same study it was observed that students living in the more homogeneous setting of college housing did experience greater reduction of the discrepancy score than did students living at home. These findings suggest that the more frequent and more realistic opportunities to test one's self against one's peers or against more concrete evidences of productivity contribute to a more realistic appraisal of the self concept. The implications for the individual's evaluation of conventional grade point achievement and its importance to him as a measure of success in college are most challenging for those who would seek to improve methods of predicting achievement.

Criterion Research Viewed in Terms of Pace and Stern's Work

A whole new area of exploration is opened up by the Pace and Stern studies (Pace and Stern, 1958; Pace, 1960) of the characteristics of college environments. These studies indicate that students feel that a particular college campus has special intellectual and social characteristics which give the institution a unique "personality"

of its own. This perception now raises the very interesting question as to whether students make their college choices in terms of their knowledge of the institution's "personality," or whether these are factors and characteristics which they discover after their associations on the campus for one or more years. Is it not likely, however, that if a student finds himself in a college environment which is not congenial to his anticipations, his achievement therein may be affected adversely? Or is it not possible contrariwise, that, finding such dissatisfaction the student will overcompensate by withdrawing from the social environment as much as possible and by enhancing his own egocentric goals through effecting the highest achievement possible? Such interesting and pertinent considerations suggest the necessity of measuring college-going students on yet another dimension. Operationally this dimension would be determined by ascertaining what characteristics in the college experience students deem important and then by correlating these characteristics with the actual experiences of students on the campus to which the high school graduate is hopefully directing himself. Here again we are dealing with noncognitive factors which unquestionably have a high relationship to success level, to perseverance, and very likely even to self concept and changes in values as a result of the educational experience.

Such research needs will take into account the dynamics of the social structure and the phenomenological field of the individual operating in that structure. It is conceivable that by considering each of these significant variables as it may influence achievement a research methodology may evolve wherein the outcomes will be amenable to a much higher level of predictive accuracy than now exists. Replicative studies or investigations concerned with prediction which do not take account of these social dynamics and their impact upon the individual's perceptions have produced varying results. A critical issue may not lie in the weakness of the instruments or of the criteria used, but rather in the fact that since the studies have been conducted within different patterns of dynamics differing outcomes have been produced.

Need for New Measures of Outcomes of Learning

Coupled with the recognition that the cultural setting in which college prediction studies are made will have important bearing upon the outcomes, is the recognition that whole new methods for

measuring these outcomes are needed. For the most part, the conventional achievement tests for measuring all of the outcomes of learning are seen to have limited scope and validity. To compare the achievement of students in one college with those of another requires that there be a prior determination that the instructional objectives and contents of instruction for the two institutions are essentially the same. Regrettably this has not been done in much of the experimentation in higher education which has made such inter-institutional comparisons. Thus, to evaluate the outcomes of college education solely or even primarily on the basis of such achievement measurement is to impose a criterion which is both artificial and inadequate.

Borrowing from the field of phenomenological measurement in psychology, new studies are developing focused on the perceptions of students with reference to certain realities rather than upon the recital of the realities themselves. When this approach is applied to achievement in the college setting, we recognize the fact that the significant determiners of behavior are the perceptions (experiences) which individuals internalize *vis-à-vis* the so-called "realities" which experimenters or observers may comment upon externally. The grade point average may be seen as the "reality," but there is already sufficient evidence to suggest that it is a less significant reality than how the student feels about his grades.

Conceptualizing Processes and Outcomes in Terms of Pepinsky's Model

In attempting to establish a fact territory upon which to base operational procedures, one is faced with a research orientation that requires properly structured hypotheses providing clues for their own testing. For testing the kinds of hypotheses which are derived in college personnel operations, Pepinsky has provided a valuable model "... for conceptualizing the processes and outcomes of student development in an educational setting" (Pepinsky, 1959).

In this model, three sets of conditions are proposed. There are first the observable antecedent conditions (the independent variables). These consist of such things as description of characteristics of the student as he enters the educational setting, the characteristics of the culture(s) in which he has developed, the characteristics of the setting and the potential interactions between the student and the setting, and other related factors. In effect, this formulation includes

(a) the careful description of the student sample being dealt with, (b) an appraisal of the way in which the subjects have developed to their present stage, and (c) a careful exposition of the setting in which the study is to be conducted.

The second set of conditions, mediating conditions, is dealt with under two headings: Observed and Inferred. These are the conditions which may now be directly seen or presumed to have certain influences upon the subjects and, in turn, may be susceptible to the impact of the subjects themselves.

The third set of data consists of the observable consequences (the dependent variable). These are the presumably measurable characteristics of the student while he is in the setting, at the time that he leaves the specific experimental setting, and on subsequent occasions and in other settings. These consequences are presumably the measurable outcomes (manifested by changes in behavior, in attitude in knowledge, or in personality structure) which have resulted from or are related to the mediating conditions which have been interposed for the student during his time of residence within the educational setting.

Careful adherence to the kind of structure which Pepinsky has proposed would insure the accuracy of controls, sufficient appraisal of the environmental dynamics to permit rigorous definition of such factors and their reproduction, and adequate description of the sample in order that the necessary replication to insure stability of findings could be made. Realistically, however, it must be recognized that the rigorous description of subjects which is required by the antecedent conditions statement is rarely, if ever, fully possible. The experimenter here is thus faced with the dilemma of choosing between doing a less than full job of such description or depending upon the techniques of random selection in order to secure his samplings. Unfortunately, in much of the research which appears in current literature, either the dilemma has not been recognized or, if it has, little effort has been made to cope with it. However excellent they may be, statistical refinements cannot fully compensate for inadequacies of basic design.

Conclusion

As experimental methods in psychology and allied sciences become increasingly sophisticated, and as the variables involved be-

come increasingly subject to evaluation, there is evidence of growing courage in the willingness to attack some of the more complex problems involving the student and his interaction with his environment. Certainly, from the evidence appearing in the current literature, it may be predicted that this direction of the search, with an awareness of the importance of the cultural dynamics, will become a chief focus of experimenters' attention in the years ahead. By the introduction of computer methods for handling larger numbers of variables and more complex research design, Fisher and Roth (1961) have predicted, "with a structured multivariate design, it may be possible for guidance and personnel workers to do more research, without holding back because of limitations of scope and resources."

If some of these new directions in both design and choice of variables become the keystones upon which a whole new attack upon the the problem of predicting college success is based, it seems likely that we shall soon be making reliable predictions of complex behavior patterns in which the grade point average will be only one of several significant criteria of success in college.

REFERENCES

- Eddy, Edward D. *The College Influence on Character*. Washington, D. C.: American Council on Education, 1959.
- Fisher, Margaret and Roth, R. M. "Structure: An Essential Framework for Research." *Personnel and Guidance Journal*, XXXIX (1961), 639-644.
- Heist, Paul and Webster, Harold. "A Research Orientation to Selection, Admission, and Differential Education." *Research on College Students* (edited by Hall T. Sprague). Western Interstate Commission for Higher Education, Boulder, Colorado, and University of California Center for Higher Education, Berkeley, California, 1960. Pp. 21-40.
- Holland, John L. and Nichols, Robert C. "The Prediction of Achievement of Different College Environments." Paper Presented at APA Convention, Philadelphia, 1963.
- Iffert, Robert E. *Retention and Withdrawal of College Students*. Washington, D. C.: U. S. Office of Education Bulletin, No. 1, 1958.
- Jacob, Philip E. *Changing Values in College*. New York: Harper and Row, 1957.
- McKee, Richard C. "An Evaluation of the Relationship between College Educational Level Achieved and Self-Concept Ratings." Unpublished Ph.D. thesis, University of Denver, 1958.
- Miller, Harold. "Factors Related to Self-Concept Changes in Col-

- lege Freshmen." Unpublished Ph.D. thesis, University of Denver, 1960.
- Pace, C. Robert. "Five College Environments." *College Board Review*, XLI (1960), 24-28.
- Pace, C. Robert and Stern, George G. "An Approach to the Measurement of Psychological Characteristics of College Environments." *Journal of Educational Psychology*, XLIX (1958), 269-277.
- Pepinsky, Harold B. "Research on the Student in His Educational Setting." *Personnel Services in Education* (edited by Nelson B. Henry). Chicago: University of Chicago Press, 1959. Chapter IX, pp. 231-246.
- Ross, G. Robert. "An Exploratory Investigation of Self-Concept Differences between Groups of College Students." Unpublished Ph.D. Thesis, University of Denver, 1955.

NON-TEST PREDICTORS OF ACADEMIC ACHIEVEMENT*

LEWIS B. MAYHEW

Stanford University

IN 1951 Benjamin S. Bloom at the University of Chicago called my superior, Paul L. Dressel, to say that he and his colleagues were on the verge of a major breakthrough in the prediction of academic success in college and asked that I be sent to Chicago the next day to discuss the matter with him. Bloom had been impressed with some preliminary results obtained from using the *Inventory of Beliefs* with University of Chicago students (Cooperative Study of Evaluation in General Education, 1953). This instrument represents a measure of authoritarianism. We had discovered that bright students who ranked at the most authoritarian end of the continuum which the *Inventory* established could not be studied in a longitudinal design simply because they had dropped out of the College at the University of Chicago during their first semester there. Although these students possessed requisite abilities to succeed, they found the discussion oriented, humanities-loaded curriculum at Chicago too threatening and consequently dropped out to attend a more orthodox institution. It has turned out that research on authoritarianism in students in general education has been of considerable significance in helping understand the dynamic of some parts of collegiate education and in learning a little more about classroom work and personality. But neither the *Inventory of Beliefs* nor the other instruments derived from it have helped the administrative applications of prediction of college success one bit. Adding a meas-

* A paper presented to the Symposium entitled "Measurement and Prediction in the College Admissions Process," Fifteenth Annual Conference on Educational Research sponsored by the California Advisory Council on Educational Research (CACER) in Los Angeles, November 15, 1963.

ure of personality to previously employed high school rank and academic aptitude measures has sometimes increased predictive power slightly, but such an addition has also resulted in actual decreases in predictive success. Since that time, many investigators have similarly searched for other non-intellective factors which might account for the considerable variance in prediction which typical correlation and regression uses of high school rank and academic aptitude revealed. The *Rorschach, Minnesota Multiphasic Personality Inventory*, and the *Taylor Scale of Manifest Anxiety* have been used in some 20 studies with a median correlation of .22 with college grade point average. Study habits tests such as the *Brown-Holtzman Survey of Study Habits and Attitudes* and interest inventories such as the *Kuder Preference Record* or the *Strong Vocational Interest Blank* have also been used—and with similar or lower results (Fishman and Pasanella, 1960).

Some Other Studies

In spite of these disappointing results—disappointing that is from the standpoint of admissions officers searching for better ways of predicting academic success—the search continues. Anne Anastasi (1960) reports in *The Validation of a Biographical Inventory as a Predictor of College Success* a study of the 1958 and 1959 classes at Fordham College. She was instrumental in leading a faculty committee to identify three groups of students: (1) the type the college wished to develop, (2) the type making satisfactory adjustment to college, and (3) those who gave evidence of maladjustment or who were all-around unsatisfactory students. The original samples of 50 students in each criterion group were used to analyze 303 scoring items on a biographical inventory with correlation coefficients consistently higher than those obtained from the College Entrance Examination Board *Scholastic Aptitude Test*—verbal and mathematics parts. Although Anastasi maintained that "... it would seem that the analysis of biographical information, which has proved fruitful in developing predictors in such areas as life insurance selling and various Air Force specialties, can be equally productive in predicting the adjustment and accomplishments of college students," her study has not yet been tested in a fully operational admissions situation.

Working along similar lines Cliff W. Wing, Jr. and Virginia Ktsanes in a study entitled "The Effect of Certain Cultural Back-

ground Factors on the Prediction of Student Grades in College" (Wing and Ktsanes, 1960) found that including information concerning social class, rural-urban origins, and type of high school attended in a prediction formula involving high school rank and academic aptitude scores could improve prediction of first semester grade point average, especially for men. The results of this study suggested that "... working class and rural men do not do as well in college as can be expected on the basis of their SAT scores and high school rank in class, and that upper-class and city men do better than can be expected on the basis of their SAT scores and high school rank" (p. 18). The study confirms the hunch that college grades are in part afflicted by problems of adjustment to college but does not seem to provide much help in establishing an overall admissions scheme. The fact that the data were significant for men but not for women underscores this point.

Paul J. Woods at Hollins College is investigating other correlates of success in college and reports some of his findings in "Correlates of Attrition and Academic Success" (Wilson, 1963). He finds that geographical region, relatives who are alumnae, parental education, desire to participate in Hollins abroad, and high school principal's estimate of ability are likely to be related to academic performance. When these factors are combined with high school rank and SAT scores he finds that he can predict failure to graduate for some categories 7 out of 10 times. The important point is the "for some groups." Thus the admissions officers are placed in an awkward position. They find, for example, that students from the South (outside of Virginia) with verbal scores under 500 are a poor risk with only 31 out of 142 graduating. However, for Virginia students and other students this particular relationship disappears.

General Evaluation

The most obvious successes of admissions officers in predicting academic performances have been with high school records or with tests on some of the various academic aptitudes. In some 263 studies high school rank correlated roughly .50 with freshman grade point averages, and in a fewer number of studies somewhat lower with post-freshman grade point averages. Correlations involving the *Scholastic Aptitude Test* of the College Examination Board, *American Council on Education Psychological Examination*, or the *Ohio State University Psychological Test* and freshman grade point av-

erage are roughly .47. Achievement tests such as the *Cooperative Tests* or the *Iowa Tests of Educational Development* showed about the same relationship. Combining high school rank and scholastic aptitude have resulted in correlations of from .37 to .83 with a median of .62. However, when more than these two intellectual indexes were employed, the gains in correlation were so slight as to be virtually useless (Fishman and Pasanella, 1960). Nor did adding non-intellectual measures assist. Although modest increases in magnitudes of correlation have been recorded in some studies, they have been too slight to suggest operational application in view of the cost of obtaining such evidence as multiple-choice version *Rorschach* scores.

The present situation then seems to be roughly as follows. Admission to college is an important problem, and it is being studied intensively. Admissions officers believe that a variety of factors should be considered in making decisions about prospective students. Past academic performance clearly is involved and has proved to be the most important single evidence upon which to base a prediction of future academic success. Academic aptitude has also proved to be of significance although both prior performance and measures of academic aptitude allow a wide variance in prediction. Motivation would seem to be an important trait, although no one apparently has succeeded in measuring it with sufficient stability to allow reasonable predictions. Although colleges try to gain some evidence concerning character, it has similarly defied quantification. Personality, special interests, or abilities, extra-curricular skills, economic status, and physical attractiveness are all judged to be of concern, but again, no operational ways of using these characteristics have as yet been uncovered. Some schools have sought to use interviews to assess such matters, but with almost completely invalid results. Meehl (1954) has shown, for example, in *Clinical vs Statistical Prediction* that a multiple correlation approach with two or three objective measures yields higher predictive values than do subject evaluations by trained people having the same data in front of them including the interview.

Need to Approach Admission Problem on an Institutional Basis

Although research on the relationship between all sorts of variables and academic achievement should be continued as important

in understanding students, it is my conviction that prediction as a basis for admissions purposes has reached a plateau and that at least for a time some other approaches should be utilized. For one thing the admissions problems of various types of institutions are so varied that generalizations have little applicability. Junior colleges with their open door policies, state institutions, prestigious private institutions, and the garden variety of private institutions each have different perplexities. Thus solutions should be approached on an institutional rather than on a national basis.

Harvard University, for example, is seeking to be a national institution having an undergraduate college population representing every region. Through use of high school rank and scores on the *Scholastic Aptitude Test* of the College Entrance Examination Board it can identify the 15 percent of a prospective freshman class that is clearly gifted and should be encouraged to seek the highest academic attainments. The same instruments will reveal the lower group which is simply unable to do work at Harvard. Then there is the large middle group from which selection must be made with the full awareness that any student rejected could probably have done as well as those who were accepted. Perhaps this point concerning probable success in the middle groups should be made explicit in an admissions policy with a genuine acceptance that chance is going to be allowed to operate. The top 15 percent might be admitted without reference to any other factors than high school rank and academic aptitude test scores. The bottom of the range might be rejected on the same ground. Then arbitrary quotas might be established for regions, for ethnic groups, or for any other features that the institution would like to see represented in its student body. Selection of people to fill these quotas then could be made by blind draws or by some other chance method. Institutions which attract a more provincial student body might allow chance alone to operate once the clearly admissible students had been procured.

Such a scheme has application for the most highly selective institutions which have the faculty and resources to stretch the abilities of even the most gifted of youth. A majority of institutions are not in such a position; yet many of them are seeking to become more highly selective as the market conditions allow. Such institutions might adopt a different approach. They might for example compute an index of academic potential consisting of rank in high school and

academic aptitude for the students doing median achievement at that institution. Then an admissions policy which would deny entrance to any student who ranked a standard deviation above or below that index of potential could be announced. If there were still too many applicants left in the available pool, some form of chance selection might be used.

State universities which have a minimum entrance requirement and a legal mandate to accept all residents of the state who meet that condition might adopt still another method. Students who meet the minimum standards for admission might be accepted according to the date of their applications. When all available spaces are filled, subsequent applicants would be accepted for the following semester, since the attrition rate can generally be known. The last applicants might have to wait until the following summer before being allowed to enter college. From what we know of the virtues of a period of out-of-school maturation this might be a sound educational procedure as well.

Junior colleges pose still another problem. In California they are required by law to accept all high school graduates as well as those (18 or over) who in the opinion of the principal could profit from post-high school education. This policy results in constantly increasing demand for junior college facilities. Yet the attrition rate between the freshman and sophomore years continues at a relatively alarming half to two thirds. This rate suggests reconsideration of admissions policies, although an open door policy can still be maintained. Is it possible to compute an index of potential for all junior college applicants based on previous academic work and aptitude test scores? Applicants whose credentials had been received within specified dates might be admitted on the basis of their relative standing and the available spaces within the institution. Those who failed to enter the first class would be admitted the second semester, again since attrition rates can be known.

Grossness of Procedures for College Admissions

This paper may sound almost anti-intellectual, since it seemingly rejects more and more precise selection devices. It is not intended to be. Rather it is an attempt to suggest the grossness of the processes of college admissions and to recommend procedures which are neither more nor less refined than the processes they are to facilitate.

Admitting college students is not unlike grading them. Professional people can do a reasonable job of grading the clearly inferior, the clearly superior, and the great middle group. Invariably when more precise distinctions than these are sought, error creeps in or the cost of making refinements reaches a point of diminishing returns. And the behavior of students in college is simply too unpredictable to warrant overly refined methods. Students drop out of college because the first week end away from home is rainy, because of an unfortunate love affair, because of a fortunate love affair, or because of a change in family financial fortunes. They succeed or fail in courses because they like or dislike a professor, because they find social life attractive or unattractive, or because their teachers are unpredictable in their grading policies. These are all human variations; and as long as they persist, one can suspect that admissions procedures will not be much more precise than they are now. Nor should they be.

Actually prediction of academic success at the level of sophistication now being practiced is relatively easy to accomplish by any of a variety of means. If an admissions officer knows students' ethnic background, home location in reference to a college or university, mother's educational level, and father's income, a *reasonable* prediction can be made as to whether the student will attend college and whether or not he will succeed. *Reasonable*, that is, as compared with predictions based upon high school grades or academic aptitude. Few, however, would support basing crucial decisions on this complex. At Stephens College we find that the third best predictor of grades (high school rank is first, and academic aptitude is second) is how students answer the question, "Do you like academic or non-academic subjects the best." Again, the dangers from using this sort of evidence administratively are obvious. Thus, colleges are left with the clearly defensible use of prior academic performance and of measured aptitude and with the considerable unaccounted for variance. This unknown variance they can handle in ways which quiet their consciences such as through use of an interview, hunches about special talents, or estimates of character which in reality are scarcely better than chance. Or they can agree that chance does play a role in life and allow it to operate within limits in college admissions. Frankly, I believe this orientation to be the healthy point of view.

REFERENCES

- Anastasi, Anne. *The Validation of a Biographical Inventory as a Predictor of College Success*. College Entrance Examination Board Research Monograph, No. 1. New York: College Entrance Examination Board, 1960.
- "Cooperative Study of Evaluation in General Education." Unpublished report. East Lansing: Michigan State University, 1953.
- Meehl, Paul E. *Clinical vs. Actuarial Prediction*. Minneapolis: University of Minnesota Press, 1954.
- Fishman, Joshua A. and Pasanella, Ann K. "College Admission-Selection Studies." *Review of Educational Research*, XXX (1960), 298-310.
- Wilson, Kenneth M. *Research Related to College Admission*. Atlanta, Georgia: Southern Regional Education Board, 1963.
- Wing, Cliff W., Jr. and Ktsanes, Virginia. *The Effect of Certain Cultural Background Factors on the Prediction of Student Grades in College*. Unpublished report. New York: College Entrance Examination Board, August, 1960.

POTENTIALITIES OF THE COMPUTER FOR MEASUREMENT AND PREDICTION WITH RESPECT TO THE COLLEGE ADMISSIONS PROCESS*

WESLEY W. WALTON
Educational Testing Service

THE advent of the computer and the present availability of large-scale electronic data processing equipment for educational purposes have brought an entirely new dimension to educational measurements and their use in predictions. This change will be felt across the entire complex that we refer to as the college admissions process. The new dimension extends from the present fact that the computer in some cases *is* and elsewhere *has* the capability of becoming a repository for a fantastically large amount of information on human performance, ability, achievement, experience, and evaluations thereof. The computer stores, retrieves, up-dates, edits, and processes data at great speed. Researchers, when they have learned how, can call up these data, treat them without taking them out of the computer, and subject them to any of the statistical techniques known to the social sciences. If we have the wisdom to ask the questions the answers to which we can use to advance measurement and prediction, the technology of computers puts at our disposal dramatically new and effective tools to bring the answers within reach.

In business, industry, and government, these tools have been found to call for new ways of doing things as well as for radical departures from conventional routines for getting jobs done. New sci-

* A paper presented to the Symposium entitled "Measurement and Prediction in the College Admissions Process," Fifteenth Annual Conference on Educational Research sponsored by the California Advisory Council on Educational Research (CACER) in Los Angeles, November 15, 1963.

ences, new inter-disciplinary amalgamations, new occupational specializations, and even new models of the scientific method have been natural outcomes of the technology of computers.

The field is growing with extreme speed. One has the feeling that educators as a group and educational researchers in particular lag far behind the rest of the pack in putting the new technology to use in the solution of important educational problems and in the conduct of educational research.

*The Computer as a Tool for Understanding
Continuity of Educational Process*

The theme of this symposium is as good a context as any other one within which to explore the characteristics of the lagging lethargy which forestalls the effective marriage of "e.d.p." and "e.p." Nay, it is perhaps the best of all possible contexts for such an exploration. Although activities related to measurements and predictions of human potential occupy younger and younger age groups (fancy the potential of my idea for a new test called TRAPT, teeth-ing ring aptitude and performance test), and although these measurements promise to come full cycle (witness the existence of the certification examinations offered by the American Board of Obstetrics and Gynecology), it is still true that the most critical stage in educational development, the stage where research can contribute most to educational improvement is represented by the transition of students between high school and college—sometimes colloquially referred to as college admissions.

One of the major limitations in the past with respect to predictive research in college admissions has been the restriction as to the number of constants and variables capable of being introduced into an experimental situation. In this new day, the sky is the limit. A second or two of computer time—or a few minutes perhaps—can be taken to test out an hypothesis, to fit certain factors into the framework, and if the fit is bad, to revise the factors or the hypothesis or both. We can test every clue and can pursue every avenue that has within it seeds which just could prove to be fertile.

With this new era already upon us, then, we can look, it seems to me, for an "explosion" of knowledge about human behavior and about the prediction of human behavior under structured sets of circumstances. Interestingly enough, the new knowledge will help

us to deal in the course of the college admissions process with the population "explosion," and the "explosion" in technology will have helped us to bring about these improved understandings of human behavior. The emphasis in this paper suggests turning the computer loose on the problem of improving prediction and on refining the "scientific" or actuarial aspect of the college admissions process.

*The Computer as a Tool to Utilization of Research
Results on a Continuing Basis*

There is a second emphasis I should like to introduce. It has to do with the utilization of research results. Traditionally, college admissions has represented a point in time and space. This has been the point at which a student is moved from secondary education into higher education. Such a dimension-less view of the educational process is one that is in this day entirely unrealistic. Each student has reached his own educational level, different from that of his colleagues at school to college transition. Each student, then, makes this transition at a unique point in his development. Education in this framework is a continuing process, and the transition between secondary and higher education should in practice be an uninterrupted and continuing flow of educational experience. It is now rarely so.

Within a framework of continuing education, we should then look at prediction as a continuing aspect of it. What we can learn to do to improve upon prediction should have equal effect both on the college admissions process and on the subsequent process through which students are placed in college class sections, are advised concerning areas of concentration and fields for major study, and are counseled with respect to educational or occupational goals.

One visualizes a long series of successive matches between individual potential and educational challenge. A student and his course would "fit" one with the other. The instruction would be within the range of the student's ability, background, and interest. Course work would be new, not a re-hash of learning already mastered in earlier years. College advising would assume an actuarial base.

As may be surmised, I am suggesting that such laudable aims might be set as legitimate objectives of educational research. If they were indeed achieved, we might have made a start at decreasing college drop-out and at enhancing retention until graduation. Pos-

sibly, new knowledge growing from expanded evaluative data and from its use in cyclical prediction could even suggest a research base for actions in curriculum revisions that the colleges might wish to take in order to keep pace with curricular and methodological improvements in secondary school programs of instruction.

Amplification of the Two Major Purposes Cited

I should like now to devote a few minutes to the further development of each of the two fundamental notions which shape the emphasis of this paper. Look with me first to the opportunity which is now ours to exploit the computer in order to advance our knowledge of the educational process. How can we find the maximum potential of the computer to the advantage of education? In a word, my answer is: not without a sound grounding of computer applications in educational research.

The computer adds new dimensions to educational measurement and to its use with respect to the college admissions process. One may cite two specific ways for experimentally using the computer to seek gains in predictive validity: (a) the introduction of new statistical techniques and (b) the expansion of criteria and predictor variables within conventional models for the study of validity. In both cases, the studies would need to be classified as applied research. The results, hopefully, would contribute to more knowledgeable admissions decisions—by the student, by his school counselor, and by his admissions dean. In both cases, access to the relevant data in computerized form opens to us new vistas pointing to marked refinements in the prediction art.

Contrasted with conventional predictive validity studies in which one or two test scores and a grade point average were fitted to a criterion datum, some fascinating furbishes to the classic design now become feasible to contemplate. Let us make the conjecture that into a student's high school record went every bit of information the school found itself willing to share with the college admissions officer. Let us say also that all these bits of information found their way into a computer and that their processing put onto magnetic tape both a transcription of the raw information and statistical representations of several sorts by which the information was reduced into summary form. Finally, let us assume that student college performance data are later edited into the tape, and that all of this

happens to the records of many graduates from many schools going to many colleges. What might be done with these data using, for example, factor analysis? Factor analyzing these data by region, by different types of colleges, or by different sizes and types of high schools, would reveal (a) the main dimensions of the data and (b) loadings on these factors in varied settings of the many and various entries in the record. There would doubtless be a grade factor, a test factor, an activities factor, and a ratings factor. Loadings at a given point in time and in a given educational setting would reveal the relative impact of bits of evaluative information on these students.

Results of such factor analyses could be used to put the schoolhouse in order—to make adjustments in order to give its several evaluative practices relatively more or less impact as circumstances might suggest. Repeated factor analyses over intervals of time could be used to determine whether corrective or modifying changes were effective as well as the degree to which they were effective.

Or look at the general question of validity. Within the current frame of reference, criteria and measurement bases might be extended both in scope and in point of time. On the criterion side, it would be interesting to look at students who entered at the beginning of the freshman year, dropped out during freshman year, re-entered at the beginning of sophomore year, or transferred at some time during first two years. It would be good to look on a vector at high school graduates who enter college indirectly from high school with work experience or with military service intervening. The list of predictor variables to be drawn upon is larger still: grade point averages, differential grade point averages, composites combining course grades and test results, composites of results from school-administered and externally-administered tests, measures of extra-curricular participation, measures of out-of-school activity, measures of personal characteristics, and assessments as to health, attendance, promptness.

These two "for instances" might serve for the moment to illustrate the point. This is the point: throughout the history of science the introduction of powerful new measuring devices or procedures has led to a great burst of research. The electron microscope expanded fields such as metallurgy; x-rays have had the same effect on genetics; and nuclear research and the consequent availability of

radioactive isotopes have revolutionized biochemistry. The computer more recently has opened up a new dimension for *research*. The new dimension is particularly pertinent to and applicable in the pursuit of *educational* research. It is incumbent upon us all to exploit the opportunity which this technology affords for the improvement of education.

Some Speculative Comments

It is a speaker's prerogative to "stick his neck out." In the last few minutes, let me finish that job.

There is no such thing today as the *point* of transition between school and college. School to college transition is better described as a *line*, and each student is at his own unique place on that continuum. Some are deeply into "college education" when they are graduated from high school; others may be college-bound but far from having finished their college-preparation when they receive their high school diplomas. Educators serious about their mission on behalf of the nation's youth can ill afford to make noises as though the student at the threshold of college educationally was born just yesterday.

I am willing to make the conjecture that the computers and the educators who learn to master them will have the capability eventually of eliminating the college admissions process. College admissions in its conventional sense is inventory taking and balance sheet making. In over-simplified form, the questions are: what does the student have to bring along with him to college and what is the match between this and what the college has for him?

When a high school reschedules students several times a year in order to adjust each student's program to his sequential levels of attainment, this sort of inventory becomes a "quarterly report," not a one time only operation. When education becomes a one to one ratio with each educational experience matched to each student's ability and readiness to assimilate it, the inventory-taking will then be even more frequent. Individual differences will have been taken account of in the fullest sense of the term.

Where does our conventional concept of college admissions fit when staff and faculty of the schools know and understand their students during high school and are enabled conveniently to communicate and to share their knowledge and understandings with

staff and faculties of colleges? Does the college then face the task of "admitting" a given student? Or rather, does it accept from the school the responsibility for the next sequences to comprise his educational experience and move him into the first of those sequences with a minimum of "wheel spinning"? Do the objectives of secondary education retain their separate nature, or do they with respect to the college-able merge with those of higher education?

Judicious use in education of the staggering capability of the computer is central to answering such questions as these.

Obviously, from an operational point of view, the answers are many moons away—decades, perhaps. From the standpoint of educational research, however, probes to find the answers should be just around the corner.



MEASUREMENT AND PREDICTION IN THE COLLEGE ADMISSIONS PROCESS: SOME POSSIBLE DIRECTIONS FOR FUTURE RESEARCH*

WILLIAM B. MICHAEL

University of California, Santa Barbara

WHENEVER one looks at articles and research reports concerning the results of predictive validity studies involving the use of standardized scholastic aptitude and achievement tests employed in the process of college admissions it becomes apparent that relatively little practical progress has been made during the past 25 years. Typically validity coefficients for single tests fall somewhere between about .30 and .50 for boys and between about .40 and .60 for girls. Rarely will a coefficient of multiple correlation associated with a combination of two or three cognitive predictors exceed .70 for either sex. This absence of apparent progress certainly has not been due to any lack of developments in statistical methodology or to any lack of advances in the capabilities of data processing equipment. Indeed, the tremendous growth of statistical methodology in the behavioral sciences has occurred at a rate far in excess of that of any organized segments or bodies of substantive knowledge in the behavioral sciences, as inspection of the December 1963 issue of the *Review of Educational Research* devoted to Statistical Methodology in Educational Research will readily reveal (Michael, 1963).

Purpose

It will be the purpose of this paper (a) to point to some possible future directions of research in the use of measurement and evaluation

* A paper presented to the Symposium entitled "Measurement and Prediction in the College Admissions Process," Fifteenth Annual Conference on Educational Research sponsored by the California Advisory Council on Educational Research (CACER) in Los Angeles, November 15, 1963.

tion procedures of potential relevance to college and university admissions and (b) to suggest possible steps largely administrative in scope that may be taken to facilitate the realization of some practical gains in the effectiveness of college admissions procedures. Although considerable, if perhaps not major, emphasis will be placed on research efforts concerning use of cognitive measures, it is the writer's contention, or bias, that future gains will result largely, though not exclusively, from intensive study of the constructs in such noncognitive areas as those associated with the dynamics of personality development and of the socialization process as they are reflected by adaptive behavior to the requirements of families of college environments. In short, from the standpoint of a theoretical orientation, consideration will be given to research potentialities in both cognitive and noncognitive domains.

The Criterion Problem

The Criterion Barrier

One reason frequently given for the absence of higher predictive validity coefficients than those usually obtained has been the lack of reliability of the criterion. In the instance of the college and university, the criterion is almost always one of grades. Under the most favorable conditions usually not more than 50 percent of the variance in the criterion has been associated with a weighted combination of predictor variables such as either a battery of aptitude and achievement tests or a composite of the high school record and one or more tests of scholastic aptitude. As is well known to the reader, the realistic proportion of variance in the criterion that may be expected to be associated with predictors used in the admission process will frequently vary between .20 and .35.

Many an admissions officer and counselor, as well as psychometrician, has pointed out to the writer that in view of the unreliability of the criterion of grades and the presence of a considerable restriction in range of talent, the proportion of predicted variance probably cannot be augmented to a great degree. The implication of such thinking is to freeze the existing criterion barrier at a point which still leaves a minimum of 50 percent of the variance—and more frequently than not 70 to 75 percent of the variance—in the criterion of grades unexplained.

These same individuals have given the impression that there is probably little that can be done about improving prediction, since professors will continue to assign grades in the same manner in the future as they have determined them in the past. An attitude of peaceful coexistence with the professors is taken for granted. In short it has been assumed that since little can be done to improve the reliability of grades, and since there probably will be an increasing restriction in range of general ability of college applicants, the admissions officer and counselor might as well settle for the status quo. In short, a quarter of the criterion loaf—25 percent of the variance accounted for—is judged to be better than no loaf at all.

Non-predicted Reliable Variance in the Criterion. In the estimation of the writer there is a considerable amount of evidence to indicate that frequently there may be a substantial proportion of reliable variance in the criterion of grades that is not predicted. One may observe that while aptitude and achievement tests often correlate about .40 to .50 with grade point averages in college the grades themselves in different college courses frequently correlate with each other between .65 and .75. It is apparent that there is a substantial amount of systematic variance in the criterion that is not being tapped by existing measures of aptitude and achievement.

For example, support for such a statement is evident from the results in two validation studies concerning the situation of student nurses in which the writer was involved (Haney, Michael, and Gershon, 1962; Michael, Haney, and Gershon, 1963). In the instance of traditional aptitude and achievement tests the predictive validities with grades in academic courses were consistent with expectations. It was noted that the intercorrelations of grades both in academic courses and in nonacademic experiences as in ward and clinical activities were much higher than the validity coefficients of the predictors with either academic or nonacademic criteria. Great care had been exercised to assure independence of judgments in the evaluation process and to minimize the introduction of possible halo effects. Clearly there were sources of systematic reliable variance in the grading and evaluation procedures which simply were not being covered by any of the predictive measures.

It was hypothesized that in addition to any possible halo effects which may not have been controlled, the systematic variance could be attributed to (a) a sort of academic savoir-faire in which there

was a perception of and empathy with a teacher's or supervisor's needs and expectations and (b) an institutional press manifested both by conformity to imposed standards of conduct, attendance, and punctuality and by practice of sanctioned and rewarded patterns of work habits. Parenthetically, it may be mentioned that efforts are now being directed toward obtaining autobiographical information from both student nurses and new applicants as well as statements and anecdotes from supervisory nurses and teachers regarding what they believe to be examples of behavior of the successful nurse. It may then be possible to formulate categories of behavior that can be evaluated for predictive purposes in the future.

Rationale for Increasing Predictive Validities. It is not a profound observation to note that the predictive validity of measures used for college admission can be raised (a) if a greater proportion of the existing reliable variance in the criterion can be duplicated by predictors than formerly and (b) if the amount of reliable variance in the criterion can be increased—a circumstance that can lead potentially to isolation of augmented amounts of common-factor variance and hence ultimately to higher predictive validity coefficients. Numerous indeed are the specific approaches that may be utilized either (a) to augment the proportion of criterion variance held in common with scores on aptitude and achievement tests or (b) to develop new criterion measures (such as improvements in the evaluation and grading systems) in which greater proportions of non-error, or true-score variance, are present.

Dependence for Gains in Predictive Validity on a Theory of Teaching and Learning

It would seem that the most promising opportunity to achieve either one of these objectives concerning the criterion, or to realize both of them, rests upon the development of a comprehensive theory of the teaching-learning process in which the basic input and output parameters can be systematically varied or manipulated and in which the corresponding interpretations of hypothesized intervening processes, or constructs, in both the cognitive and noncognitive domains can be suitably rendered or appropriately modified within the framework of such important environmental characteristics as the nature of the curricular objectives and the specificity of the institutional context—that is, the academic climate of the classroom.

department, college, or university. A very clear and detailed description of a unified theory of teaching-learning behavior based upon an *information-system model* in which, among other things, the effect on teacher behavior of student interaction and feedback is emphasized, has been set forth by Ryans (1963).

In his very significant chapter Ryans also summarized important conceptualizations of three other investigators whose thinking displayed many points in common with his own model: (a) Smith's (1960) perception-diagnosis theory of teaching in which a cycling process of the perceptions by students and teachers of the expectations of the others in the instructional process takes place, (b) Smith and Meux's (1962) formulation of a theory of teaching as a system of social action in which strategies and logical operations are divided toward attainment of goals concerning experiences with subject matter and toward manipulation of the subject matter; and (c) Turner and Fattu's (1960) model in which the teacher is viewed as a problem-solver and decision-maker. One of the key values underlying these heuristic theoretical formulations is that both cognitive and noncognitive processes of teachers and students, as well as their interactions in the social framework of the learning situation, may be hypothesized and described—at least tentatively—in the broad context of the internal and external environment of the student and the teacher.

To facilitate the understanding of such comprehensive systems, miniature theories can be built into the overall framework to help explain portions of the domain of the learning-teaching process. For example, the structure-of-intellect model proposed by Guilford (Guilford and Merrifield, 1960; Guilford, 1962), though hardly a miniature theory, can become an integral part of the very comprehensive framework of Ryans' information-system model. Moreover, the highly useful and communicative paradigms or schemata that have been proposed by Gage (1963) in his chapter in the *Handbook of Research in Teaching* edited by Gage (1963) himself can be used to great advantage to portray many of the essential components and interrelationships such as the internal inputs and external inputs, the information-processing itself, the outputs of information-processing, and the products of teacher behaviors, which themselves are student behaviors.

An indicated previously, the distinct value of a theoretical frame-

work is to guide research concerning the basic nature of the teaching and learning process in a variety of contexts. In light of research findings the theory can be modified many times, and new hypotheses can be developed continually and subjected to verification or refutation. As knowledge is gained concerning the characteristics of intellectual and nonintellectual processes—especially the interaction present in the social processes of teachers, students, and student groups in institutions of higher learning with similar value systems—the criteria of academic success should be described more accurately and more completely than they have been in the past, and corresponding measures of those more adequately described characteristics simultaneously can be developed and validated. Indeed, progress in augmenting predictive validities in the past has been slow, and progress of the future cannot be expected to be rapid.

Without a comprehensive theoretical framework it would appear, however, that at best only limited empirical achievements of a transitory value will be realized—achievements which may be expected to collapse repeatedly when crossvalidation studies are undertaken. Once the constructs in the criterion measures of the teaching and learning process that are associated both with the interactive cognitive and noncognitive behavior of students and teachers and with the objectives of the curriculum in a given institutional context can be validated in replicated studies, gains in prediction of college success should be forthcoming from those predictor measures which represent an operational translation of the same constructs in the various criteria of evaluation of college performance.

Implementation of a Prediction Program Anchored to a Theoretical Structure

Before any specific types of promising efforts aimed at the improvement of the predictive validity of measures utilized in college admissions work are cited, it would seem important to stress that to be successful, research endeavors should reflect the participation of students, faculty, and administrators as well as of the research specialists in measurement and their assistants. Unless students and faculty, in particular, feel some degree of personal involvement in a comprehensive program aimed at the improvement of selection procedures in the college admissions process and in the counseling experiences that follow throughout the years of college attendance, it

is doubtful that any substantial progress in raising the predictive validity of selection devices will occur.

Need for Institutional Research Centers

From an operational or administrative standpoint, it would appear that each college or university would need to have a center for institutional research that unlike many in existence is staffed by adequately trained and professionally competent individuals who themselves also do some teaching as well as research and administrative work. Moreover, there should be some sort of rotational plan in which members from various academic departments become involved from time to time in committees with both an advisory and participating function. Such participation seems to be lacking in most institutional research centers. Cooperation for undertaking novel research studies is much more probable when faculty members have a voice in the affairs of a research center than when they do not.

Risk of a Hawthorne Effect. As students and faculty become involved in research activities pertaining to the evaluation of the college program, the nature of the criterion such as the determinants underlying the assignment of grades may be expected to change. With such cooperation as that afforded by an institutional research center there is admittedly the risk of a form of criterion contamination that may occur as a consequence of a possible Hawthorne effect associated with student-teacher involvement in a college-wide research project.

From the standpoint of motivation professors may be expected to put forth greater efforts than is customary in defining their course objectives, in building examinations centered about these objectives, in encouraging an increasing amount of student participation in the teaching-learning process, and in trying to win a position of prominence for their classes. Likewise, students who in any way have been consulted about their participation in experimental studies may be spurred on to increase their efforts and to go out of their way to please their professors and to bring distinction to their major department or to their particular class. In crossvalidation studies with new groups of students and faculty who were not similarly involved in an emotional identification with the project, the amounts of shrinkage of predictive validity could be shocking indeed.

A case in point arose a number of years ago in the English department at a neighboring university. There had been a department-wide committee which with the cooperation of the office of the university examiner carried out a comprehensive and coordinated program of testing and evaluation in freshman English. The entire departmental faculty formulated common course objectives, participated in building common examinations for all sections, agreed in advance on keys for objective tests and on criteria for assignment of points to essay examinations, and took part both in scoring the essay examinations and in grading term papers as well as in assigning grades. Although an estimate of the reliability of the grades for two semesters was not obtained, it could easily have been in excess of .80. Predictive validities of scholastic aptitude tests administered to freshmen fell between .60 and .70 relative to a criterion of grades in the two semesters of English.

This pleasant state of affairs continued for about five semesters until the departmental chairman went on a sabbatical leave. His replacement who encouraged a completely laissez-faire policy for professors in his department urged each of the instructors in the freshman program to organize his own course, to make his own type of assignments, to pursue any evaluation technique that seemed appropriate, and to use his own conscience in assigning grades. At the end of the next semester, the validity coefficients dropped from magnitude near .65 to values in the vicinity of about .20.

Needless to say, reports reached the testing officer concerning a large number of students who were trying to enroll in courses of those particular instructors who not only gave rather lenient marks, but also used examination and evaluation procedures compatible with the expectations or preferences of particular students. Under such circumstances it would appear feasible to believe that there was a marked change in the factorial composition of the variance of the criterion. One might hypothesize that originally a high proportion of the reliable variance in the criterion was associated with clearly stated and carefully formulated instructional objectives that were consistently evaluated on examinations. In the latter situation associated with the laissez-faire policy, there was apparently not only a drop in the amount of reliable variance in the criterion measures, but also a higher proportion of the variance in common with non-cognitive constructs such as those involving a sensitivity of percep-

tion in students of the instructors' expectations about students and of the instructors' own psychological needs—a social shrewdness in the academic situation necessary for obtaining high grades. It may well be that in the relatively unstructured learning situation the proportion of variance associated with noncognitive constructs such as the interpersonal perception of students and instructors is great relative to what it would be in a highly structured learning environment in which course objectives and requirements are clearly set forth and are systematically and objectively evaluated.

Benefits from an Institutional Research Center

Even if the previously mentioned Hawthorne type of effect may take place in coordinated research programs concerned with the improvement of college success, it may well be, according to certain value criteria, that over a prolonged period of time essentially permanent changes of a positive and beneficial nature in the teaching-learning processes will take place. Certainly clarification of the objectives of courses and improvement of examination procedures should serve to enlarge the relative proportions of reliable criterion variance that can be predicted from standardized aptitude and achievement tests in the cognitive domain. If in addition certain important objectives associated with somewhat more intangible objectives such as appreciations, esthetic values, and social values of the educational process can be set forth in relatively clear terms, it may well be possible through the cooperative efforts of students and faculty to develop reliable measures of such characteristics that could be used both for evaluative purposes and in predictive studies.

An additional advantage of the research center would be that of a channel for communication between faculty of different departments or divisions, between faculty and administrators, and between faculty and certain dedicated alumni who are interested in improving the educational program of the college or university. In any attempt to promote a particular academic philosophy or policy as might be the case in developing a campus climate or a college image, the evaluation of such an activity could profitably be directed and coordinated by the staff of the research center. In any regional or national research project concerned with college admissions and testing programs, the center would be the logical administrative and coordinating unit in the college or university.

*Some Suggested Research Approaches to the Improvement of
Prediction of College Success*

Contributions Primarily in the Cognitive Domain

Review of Guilford's Structure-of-Intellect Model. Within any broad theoretical framework that may aid one to understand the teaching-learning process, there are several specific avenues of research pertaining to measurement and prediction which may be suggested. In the cognitive domain the whole area of creative abilities, for example, affords a tremendous challenge, especially since a theoretical model in terms of Guilford's structure of intellect already exists (Guilford, 1962; Guilford, Fruchter, and Kelley, 1959; Guilford and Merrifield, 1960). Guilford's three-way grid, or cube, which portrays the three dimensions of (a) contents (resembling inputs), (b) operations (denoting process constructs), and (c) products (outputs in various types of units) is a helpful model that can be advantageously employed as a theory when the relationships between its constructs and the world of empirical observation are hypothesized.

As the reader may recall from his study of previously cited writings by Guilford, the operation called divergent production has been shown to be related to what are judged to be a number of creative measures. In the domain of divergent production, there is considerable emphasis placed on a variety of output, or on the generation of new information, from the same source of given information as portrayed in open-ended test items. In other words, the individual who comes up with clever, novel ideas and who exhibits an abundance of ideas per unit of time—an inferred fluency ability—is frequently judged to be creative. There is also the creative individual who is disposed to seek a variety of answers without being hindered by a previously well established mental set—a form of behavior which is considered to be flexible and adaptive. Moreover an individual is often said to be creative if he can be evaluative in his output—that is, sensitive to problems and critical in judgment concerning the merits of various products (outputs).

Divergent versus Convergent Thinkers. There is a serious doubt in the writer's mind concerning how many professors in college actually give encouragement to the student who often comes up with novel, clever, original, or even bizarre ideas. It does seem safe, however, to hypothesize, as Guilford has done, that many individuals

tend to be divergent thinkers rather than convergent thinkers. The latter group consists of persons who attempt to achieve a particular, conventionally accepted, or clearly designated form of behavior from information that has been specified or formally given as in the instance of multiple choice items. As may be well known to the reader, the traditional tests of scholastic aptitude and achievement tend to stress convergent thinking processes, or simply convergent production.

Implications for Admissions. From a practical admissions standpoint, there is obviously a need to identify and to select those students in college who may be somewhat marginal in measures of convergent production, but who with high standing in abilities of divergent production can achieve to a high degree—especially in courses in which professors encourage somewhat radical departures from conventional procedures or assignments. For example, in the arts and in the laboratories of the sciences, measures of divergent production would seem to hold great promise of yielding a substantial degree of predictive validity—especially if the instructors of such courses encourage freedom of thought and freedom of action on the part of their students.

Implications of the Specificity of the Instructional-Learning Process. From the work which Hills (1955) undertook a number of years ago, there is evidence concerning the utilization of creative talents by upper-division and graduate students majoring in mathematics at three different institutions. His findings revealed a marked specificity of the presence of creative abilities depending upon the particular curriculum-criterion-institution context. As one might expect, in each of two institutions in which utilization of a particular type of an hypothesized ability that was intended to represent a creative approach to solving problems was encouraged, that type of ability appeared in his factor analysis. In the institution in which little attention was given to an approach in teaching that encouraged students to be creative relative to factors hypothesized, only the more traditional abilities in convergent production emerged in the factor analysis. The moral to this investigation is that the existing variance in the criterion cannot be changed to reflect creativity unless the instructional-learning process itself is correspondingly altered.

Implications for Prediction of Success in Graduate Work. There are additional implications of this so-called creativity approach for

graduate study. Although the restriction of range in talent present for graduate students will serve to attenuate the predictive validity of aptitude measures and although the criterion of graduate grades is frequently dichotomous (A or B grades being assigned) or at best trichotomous, there is certainly a need to exert concerted efforts to improve the degree of predictive validity of devices used for graduate selection—again through study of the criterion variables. It would seem that particularly in the graduate program, in which original research efforts and independent contributions of students are supposedly encouraged and emphasized, there is a great challenge for professors to work cooperatively with an agency such as a research center to develop diversified criterion measures that reflect constructs of creative ability. As mentioned previously, once such constructs can be isolated and the implications of their meanings incorporated into the instructional procedures, it should not be too difficult to develop standardized predictive measures to duplicate the same constructs—tests that could be used in part for purposes of admission of graduate students.

Implications for the Role of the Institutional Research Center. The existence of a coordinated program at a research center would facilitate the design and completion of experimental studies in which creative experiences are an integral part of the curriculum. The research center could assist in the development and use of predictive and evaluative devices in these programs—tests and scales which even the more conservative professors might feel that they could accept. In light of such efforts, constructs in the course criteria might well be isolated, identified, and subsequently translated into achievement test measures. If such an objective could be realized, the amount of variance of a reliable nature in the criterion of grades could be augmented considerably. If the benefits of such a program, however, were to persist, the entire educational project would need to be tried over a period of several years.

Consideration of Bloom's Taxonomy. One final suggestion for a research effort in the cognitive domain is concerned with the comprehensive taxonomy of educational objectives prepared by Bloom and others (1954). The *a priori* constructs which were derived from the context of achievement test items might well be related in a systematic way to Guilford's structure-of-intellect model—a suggestion made by the writer several years ago (Michael, 1957, 1961). In a study with children, Schmadel (1960) demonstrated that criteria

tion measures of complex achievement tasks which stressed use of the higher level cognitive processes of synthesis and evaluation in the taxonomy could be predicted more validly from those measures of hypothesized creative thinking abilities in Guilford's structure of intellect concerned with sensitivity to problems, conceptual foresight, ideational fluency, and originality than from measures of traditional convergent production found in standardized achievement and general intelligence tests. It is regrettable that systematic efforts at the college level have not been expended along the lines pursued by Schmadel in order that the degree of correspondence between the constructs of Guilford's model and those in the taxonomy could be ascertained.

Contributions Primarily in the Noncognitive Domain

Fishman's Social Psychological Theory. In his highly cogent discussion of the role of nonintellective factors in college selection and guidance research, Fishman (1962) pointed out from a survey of 580 studies in college guidance and selection conducted during the decade from 1948 to 1958 that the average multiple correlation between the usual predictors (high school grades and scores in standardized tests of scholastic aptitude) and a criterion of grade point average in college was about .55 and that the gain in multiple correlation associated with the addition of a personality test score to one or two cognitive predictors with the criterion held constant was customarily less than .05. He emphasized that personality tests and other noncognitive measures could conceivably contribute more than they have contributed if they were actually measuring something sufficiently dissimilar to that found in the usual predictors—even when allowance is made for the unreliability of noncognitive measures as well as for the unreliability of the criterion variables. Subsequent to delineating the difficulties involved with both nonintellective criteria and nonintellective predictors and after presenting reasons why so few colleges have carried out, or can be expected to carry out, predictive studies involving nonintellective criteria in light of their emphasis on academic goals, Fishman concluded on theoretical grounds that he must remain pessimistic concerning realization of long-range gains from current approaches. He argued that differential weighting systems of existing cognitive predictors which correct for differences in the academic and societal characteristics of high schools, of the families, and of communities of origin of the

college applicants will increasingly erode any contributions that nonintellective measures can make to the prediction of intellective criteria. Moreover, he noted that high school grade point averages themselves reflect substantial sources of variance in nonintellective factors.

In view of his many arguments, he proposed an alternative approach to the familiar personnel selection device—an approach that consisted of several environmental models in a social psychological theory for understanding the behavior of students. His models represent various simultaneous permutations in the presence of sameness or differences in the individual when he is in college or when he was in high school, of sameness in high school and college environments, and of differences in high school and college environments—environmental differences in turn which may be classified as constant or variable. The differences in the individual in college that differentiate him from what he was in high school may be attributed to a developmental change or to a random change.

In considerable detail Fishman explained the implications of the nine resulting models for uses of various combinations of individual and institutional intellective and nonintellective predictors. He considered the place of individual contingency moderators which intrude between predictor and criterion variables in unforeseen ways and also gave attention to the presence of complex interactions among the variables he postulated. These exciting models should serve as fruitful guide lines in the planning of systematic research studies and as safeguards for preventing the execution of relatively sterile investigations or the replication of essentially pointless investigations. These models thus afford a means not only for augmenting the amount of reliable criterion variance that may be isolated in both intellective and nonintellective criteria, but also for increasing the proportion of identifiable and measurable common-factor variance in a variety of criterion measures.

Work of Stern and Pace concerning College Environments. Clearly related to Fishman's proposed approach have been the efforts of Stern (Stern, 1962, 1963; Stern, Stein, and Bloom, 1956) whose concern for college environments in relation to the learning process suggests ways for matching students to the particular sort of college environment that will meet their psychological needs and expectations and which indirectly also will be harmonious with the perceptions of their fellow students, the faculty, and the adminis-

trators. Similarly, the significant work of Pace and Stern (1958) who have approached the problem of measurement of college environment is also directly pertinent. Through the development of his test called CUES, Pace (1962) has succeeded in furnishing measures on five factors of college climate: (a) *practicality*—a practical-instrumental orientation in contrast to an emphasis on abstract and theoretical matters on campus; (b) *community*—a friendly, cohesive group-oriented campus with a supportive and sympathetic atmosphere in contrast to one of personal autonomy and cool detachment; (c) *awareness*—an emphasis on self-identity and self-understanding and on an idealistic and esthetic concern for the condition of man in contrast to a lessened interest in soul searching; (d) *propriety*—a polite and considerate environment in contrast to an assertive, demonstrative, free-wheeling, individualistic one; and (e) *scholarship*—a highly competitive scholastic and academic environment in contrast to a lessened emphasis on intellectual endeavor for its own sake. These scales along with carefully prepared biographical data inventories which are geared to the requirements of a theoretical framework such as the one set forth by Fishman may add substantially to the prediction of college success within the specificity of a given institutional context.

Interpersonal Perception of Faculty and Students. Again within the sort of theoretical structure furnished by Fishman there is an opportunity to exploit dimensions of interpersonal perception held by faculty and students concerning their respective intellectual and social goals. There is also the possibility of developing items in some type of inventory which will pick out the student who is especially receptive to the particular requirements that individual professors set for the awarding of high marks. An examination of both Fishman's model and of the contents of social behavior to be found in Guilford's structure of intellect could be helpful. The Guilford model provides for nonverbal characteristics involved in human interactions in which an awareness of the thoughts, needs, intentions, and attitudes of other persons and of one's self may afford an opportunity for defining what the writer believes to be an academic savoir-faire factor. The student who is high in this dimension can quickly adapt to changing environmental requirements within one classroom or from classroom to classroom within a given institution despite the presence of any relatively consistent pattern, or communality, of environmental influences making up the campus climate.

This factor of academic shrewdness may not be too unlike several of the behaviors suggested in a recent article in *Time* (Time, 1963) in which students employing what might be called behavioral manifestations of a psychopathic deviant syndrome attempt to "con" their professors into giving them high marks. For example, a perceptive student may keep abreast of the most recent articles that a professor has written, may ask questions calculated to please the professor, may sit toward the front of the room and look admiringly at the professor—eye to eye, may wear previously earned or borrowed scholastic honor keys at strategic times, may make sure that the professor learns his name early in the course, may assume a contrary position on an issue and most respectfully and tactfully disagree with the professor in order to arouse the professor's interest, may flatter the professor in highly subtle and non-obvious ways as in taping his lecture or in asking him to speak to a student group, may go out of his way to have fellow students tell the professor how intellectually proficient their friend (the student) is, may make an appointment to show the professor an outline for the assigned term paper in order to insert the professor's ideas into the paper, may systematically gather samples of previous examinations and term papers in order to meet the expectations of the professors, may offer at a strategic moment to assist a professor with his research or laboratory work, may make use of her sex appeal and personal charm when it is deemed appropriate, may play on the sympathies of professors in a variety of ways, and may exploit any opening that allows the furthering of an impression that he is a dedicated student—a person deeply interested in higher learning, in knowledge for its own sake, and in the early publication of the truly significant work which his esteemed professor is doing to win the reprint race.

Summary

An attempt has been made to suggest ways in which the predictive validity of measuring devices employed in the process of college admissions may be enhanced. For the most part, these suggestions have been related to how a theoretical structure in both the cognitive and noncognitive domains of behavior may serve to guide research efforts toward the understanding of the instructional-learning process, toward the validation of important constructs isolated by research studies anchored to a theoretical framework, and toward the development of new measures which may represent these

constructs and thus may serve to predict with increased accuracy both existing and new sources of reliable variance in admittedly factorially complex criteria. For the implementation of these research efforts, it has been suggested that each college or university establish a center for institutional research that is competently staffed. Such a center will afford a means for the completion of coordinated and systematically directed research with which not only the specialists at the center, but also faculty and students may closely identify themselves. It is the writer's conviction that only through use of theoretically based procedures that involve all groups of the college community will any substantial future gains be made in the level of predictive validity of both cognitive and noncognitive measures used for purposes of college admissions and placement.

REFERENCES

- Bloom, Benjamin S. and others. *Taxonomy of Educational Objectives* (Preliminary edition). New York: Longmans, Green, and Co., 1954.
- Fishman, Joshua A. "Some Social-Psychological Theory for Selecting and Guiding College Students." In Sanford, Nevitt (Editor). *The American College: A Psychological and Social Interpretation of the Higher Learning*. New York: John Wiley & Sons, 1962.
- Gage, N. L. (Editor). *Handbook of Research on Teaching*. Chicago: Rand McNally & Co., 1963.
- Gage, N. L. "Paradigms for Research on Teaching." In Gage, N. L. (Editor). *Handbook of Research on Teaching*. Chicago: Rand McNally & Co., 1963.
- Guilford, J. P. "What To Do About Creativity in Education." *Eleventh Annual Western Regional Conference on Testing Problems*. Los Angeles: Educational Testing Service, 1962.
- Guilford, J. P., Fruchter, Benjamin, and Kelley, H. Paul. "Development and Applications of Tests of Intellectual and Special Aptitudes." *Review of Educational Research*, XXIX (1959), 26-41.
- Guilford, J. P. and Merrifield, Philip R. *The Structure of Intellect Model: Its Uses and Implications*. Reports from the Psychological Laboratory, No. 24. Los Angeles: University of Southern California, 1960.
- Haney, Russell, Michael, William B., and Gershon, Arthur. "Achievement, Aptitude, and Personality Measures as Predictors of Success in Nursing Training." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 389-392.
- Hills, John R. *The Relationship between Certain Factor-Analyzed Abilities and Success in College Mathematics*. Report from the Psychological Laboratory, No. 15. Los Angeles: University of Southern California, 1955.

- Michael, William B. "Differential Testing of High-Level Personnel." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVII (1957), 475-490.
- Michael, William B. "Problems of Validity for Achievement Tests." In Huddleston, Edith M. (Editor). *Eighteenth Yearbook of the National Council on Measurement in Education*. Ames, Iowa: the Council, 1961.
- Michael, William B., issue editor. "Statistical Methodology in Educational Research." *Review of Educational Research*, XXXIII (1963) 451-586.
- Michael, William B., Haney, Russell, and Gershon, Arthur. "Intellective and Non-Intellective Predictors of Success in Nursing Training." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIII (1963), 817-821.
- Pace, C. Robert. *CUES: College and University Environmental Scales*. Preliminary Manual: Description, Norms, Uses. Princeton, N. J.: Educational Testing Service, 1962.
- Pace, C. Robert and Stern, George G. "An Approach to the Measurement of Psychological Characteristics of College Environments." *Journal of Educational Psychology*, XLIX (1958), 269-277.
- Ryans, David G. "Assessment of Teacher Behavior and Instruction." *Review of Educational Research*, XXXIII (1963), 415-441.
- Schmadel, Elnora. *The Relationship of Creative Thinking Abilities to School Achievement*. Unpublished doctoral dissertation, University of Southern California, 1960.
- Smith, B. Othanel. "A Concept of Teaching." *Teachers College Record*, LXI (1960), 229-241.
- Smith, B. Othanel and Meux, Milton O. *A Study of the Logic of Teaching*. U. S. Department of Health, Education, and Welfare, Office of Education, Cooperative Research Project No. 258. St. Louis: Graduate Institute of Education, Washington University, 1962.
- Stern, George G. "Environments for Learning." In Sanford, Nevitt (Editor). *The American College: A Psychological and Social Interpretation of the Higher Learning*. New York: John Wiley & Sons, 1962.
- Stern, George G. "Characteristics of the Intellectual Climate in College Environments." *Harvard Educational Review*, XXXIII (1963), 5-41.
- Stern, George G., Stein, Morris I., and Bloom, Benjamin S. *Methods in Personality Assessment*. Glencoe, Ill.: Free Press, 1956.
- Time, "Students, Conning the Professor," *Time* LXXXII (November 1, 1963), 56, 58.
- Turner, Richard L. and Fattu, Nicholas A. "Skill in Teaching, A Reappraisal of the Concepts and Strategies in Teacher Effectiveness Research." *Bulletin of the School of Education* (Indiana University), XXXVI (1960), 1-40.

THE FACTORIAL COMPOSITION OF AGCT "SUBTESTS" ALONG WITH COLLEGE APTITUDE ITEMS AND HIGH SCHOOL GRADES¹

LEROY WOLINS²

Iowa State University

AND

ROBERT PERLOFF²

Purdue University

THE results of this study suggest a possibly useful guide for constructing items designed to measure reading comprehension and ability to interpret charts and graphs. In this study item clusters refer to a given reading passage, to a set of data pictorially represented, and to other well known factors such as verbal, quantitative, and spatial aptitudes. These groups of items were factor analyzed along with high school grades.

The sample consists of 238 male high school seniors. Both SRA Form 4 of the Selective Service College Qualification Test (SSCQT) and Form 1c of The Army General Classification Test (AGCT) were administered to these students. The mean AGCT standard score for this sample is 110.3 with a standard deviation of 13.7 standard score points.

Table 1 identifies the variables and their factor loadings. The first 19 variables are item types from the Selective Service College Qualification Test. Variable 20 is 6th semester high school class rank. (In reference to the rather high loading of .42 on the Knowledge of Arithmetical Operations factor, one should note that many

¹ Adapted from a paper presented by the investigators at the 64th Annual Convention of the American Psychological Association, Chicago, 1956.

² The opinions expressed herein are those of the authors and do not necessarily reflect official views or policies of the Selective Service System.

TABLE 1
Factor Loadings (decimals omitted)

Item type	(Number of items)	General Intelligence	Vocabulary	Reading	Data Interpretation	Quantitative Reasoning	Perceptual Speed	Arithmetical Operations	<i>h</i> ²
1. Antonyms	(10)	(41)	(52)	00	04	-04	06	13	48
2. Synonyms	(9)	(29)	(46)	05	-02	00	-04	-10	31
3. Verbal Analogies	(10)	(48)	(26)	00	-01	06	01	-05	30
4. Sentence Completions	(9)	(37)	(45)	(20)	12	02	02	-10	40
5. Reading Passage	(6)	(44)	(40)	(22)	03	00	10	-09	42
6. " "	(6)	(53)	(22)	(45)	06	12	-11	10	57
7. " "	(7)	(45)	05	11	-03	08	09	-11	24
8. " "	(6)	(44)	(31)	(25)	06	-10	05	06	37
9. " "	(6)	(40)	(32)	(24)	-01	05	-09	08	34
10. " "	(6)	(35)	(23)	(26)	00	00	-11	13	27
11. Data Interpretation	(6)	(55)	-04	-04	-08	00	01	-02	31
12. " "	(4)	(58)	-05	-12	(20)	01	10	-13	42
13. " "	(3)	(48)	03	01	(18)	05	05	05	27
14. " "	(6)	(54)	-03	-09	(26)	-04	-04	12	39
15. " "	(6)	(59)	00	-01	(22)	-01	-06	-05	40
16. " "	(4)	(34)	00	13	(42)	07	04	10	33
17. " "	(4)	(40)	07	05	(37)	04	-05	14	33
18. " "	(4)	(29)	08	-06	(23)	-02	08	08	16
19. Arithmetic ¹	(38)	(63)	05	00	-02	(38)	07	(55)	85
20. Sixth Semester Class Rank ²		(55)	(30)	00	-01	(15)	-01	(42)	59
21. AGCT Verbal (early items)	(6)	(28)	(22)	-10	10	-03	10	-14	18
22. AGCT Arithmetic (early items)	(6)	(29)	-10	03	00	(45)	14	00	32
23. AGCT Space (early items)	(6)	12	05	-02	05	11	(20)	01	07
24. AGCT Verbal (late items)	(6)	(26)	(42)	-05	10	-03	13	-10	28
25. AGCT Arithmetic (late items)	(6)	(45)	10	02	-04	(35)	(37)	01	47
26. AGCT Space (late items)	(6)	(41)	-11	11	00	-09	(56)	10	52

¹ The Mathematics section of the SSCQT includes problems using numbers and symbols as well as word problems similar to those found in the AGCT.

² In this school College Algebra was frequently taught during the sixth semester.

of these students took college algebra that semester.) Variables 21-26 are sub-tests from the Army General Classification Test. Variables 21-23 are groups of 6 items occurring early in this speeded test and almost all the students attempted them. Variables 24-26 are groups of 6 items occurring later in the test and they were not attempted by many of the students.

Method

These 26 variables were intercorrelated and factor analyzed by means of Thurstone's (1947) group centroid technique. A general factor was introduced to attain simple structure along with an orthogonal solution. The factors were rotated for meaningfulness graphically and the communalities were adjusted by Wherry's (1949) iterative technique. The intercorrelations and residuals are shown in Table 2.

Results

The results of the factor analysis are reasonably clear except possibly for the interpretation of the Reading and Data Interpretation factors. The meaning of these two factors can be more clearly understood by contrasting the apparent content of the variables loading upon them the greatest and the least.

Variable 6, a reading passage dealing with the origin of comets, is loaded highest on the reading factor. This passage contains unusual words such as "retrograde," but the meaning of the word is made clear by examining the way in which it is used. Such words are *not* formally defined in this reading passage. The questions refer *indirectly* to the reading passage and content. They do *not* ask for facts given explicitly in the passage, but deal rather with what is implied. In no case do the questions require what would be considered uncommon knowledge. The item which best characterizes this passage is:

"The best title for this passage is___?" [Followed by (A), (B), (C), (D), (E)].

Variable 7, a reading passage describing curling, is made up of sentences such as the following: "Curling is played on an ice rink by two teams, each having four players, and uses circular stones, two for each player." Another sentence is: "At each end is a target called the *tee*, around which there are four concentric circles forming the

TABLE 2

Intercorrelations (above diagonal)—Residuals (below diagonal)

1.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
2.	02	1	36	33	30	43	34	32	30	21	24	23	21	24	15	16	17	19	32	45	20	08	06	30	27	20
3.	00	00	1	25	38	31	26	28	22	10	08	14	15	19	07	13	14	12	16	23	27	07	00	31	15	04
4.	-08	05	-02	03	28	35	33	26	25	22	27	29	22	24	13	27	25	15	37	32	24	12	16	28	31	13
5.	05	-02	03	06	46	37	23	37	33	24	13	21	23	22	24	27	19	16	24	26	31	05	16	32	29	12
6.	00	-01	01	-02	04	45	27	39	33	28	15	17	21	16	25	22	28	14	35	31	12	09	17	23	23	20
7.	-02	-03	09	01	01	05	34	42	43	33	23	29	33	23	29	29	22	15	43	46	17	21	08	20	27	20
8.	00	03	-02	02	01	01	01	24	19	13	25	28	24	21	27	11	11	14	30	16	13	18	12	24	20	26
9.	-01	01	-02	00	-01	02	-02	03	36	35	19	21	24	20	26	22	28	14	27	33	14	07	04	25	23	24
10.	03	01	00	-03	00	-05	-04	06	07	12	09	16	20	16	10	10	17	17	30	39	17	09	-02	26	23	13
11.	01	-04	00	-04	-07	-02	01	-02	-03	-04	34	24	30	32	08	27	26	24	31	25	16	18	09	17	24	18
12.	03	-06	00	00	-07	06	02	-01	-03	-04	03	38	32	34	27	28	26	19	33	23	21	26	01	17	28	24
13.	00	00	05	02	-03	05	03	01	06	-01	00	07	35	35	28	30	32	26	36	35	09	21	10	19	25	20
14.	-02	02	-02	03	-03	-04	01	-03	-03	02	-04	03	34	30	33	33	24	17	31	24	17	17	07	21	23	23
15.	00	02	-04	00	-01	-03	02	00	00	-04	01	05	04	-03	33	32	10	28	21	24	24	18	03	18	24	23
16.	-01	-02	-03	08	04	01	-04	01	04	-06	-07	01	03	01	05	32	32	30	32	21	14	10	06	08	23	14
17.	04	01	05	-03	07	-08	-05	-04	-04	02	-00	-01	-01	-00	-07	01	30	30	32	35	12	16	16	14	18	19
18.	01	03	00	02	-01	00	03	-02	05	05	-09	03	00	02	-04	-10	09	01	23	27	16	04	05	19	18	15
19.	-03	01	05	-03	09	-01	04	-02	-03	-01	-03	03	04	-02	-03	-01	-02	01	61	01	01	32	16	15	54	30
20.	02	-03	-01	-03	-01	04	-06	-01	04	09	-03	-02	00	02	-06	-02	05	07	-04	-04	11	22	07	19	33	26
21.	-03	08	04	-10	-10	04	-01	-03	03	-04	02	00	-07	01	05	03	-02	04	-10	-04	-01	05	05	28	21	09
22.	02	04	-02	-03	-02	03	00	01	-01	07	01	07	05	04	02	-04	04	-04	-04	03	-01	10	10	01	34	10
23.	-03	-05	08	08	08	02	05	-03	-07	00	03	-09	01	10	-04	-02	09	-02	02	-03	-02	-01	07	07	14	20
24.	-03	04	04	01	-07	01	09	00	05	-08	05	-01	02	-07	01	-04	-01	06	02	-03	04	-03	-01	03	22	14
25.	02	-01	04	07	-05	00	-07	01	03	-02	-01	-01	00	01	00	05	01	03	08	00	04	01	-03	03	22	14
26.	04	00	-03	00	-01	02	04	03	02	08	-05	-03	-02	02	02	-04	05	-01	-02	05	-03	-07	05	02	03	39

Note: Decimals omitted in this table.

house." This reading passage is merely the description of the game of curling. It does involve formal definitions. The items referring to this passage deal with things explicitly stated in the reading passage or with things that require outside information. An example of the latter is the following:

"The method of scoring in curling resembles most closely that of

- (A) shuffleboard
- (B) badminton
- (C) bowling
- (D) ice hockey
- (E) cricket."

Variable 16, the variable loaded highest on the Data Interpretation factor, is a table rather than a graph, but this does not appear to be the significant determiner of its high loading since the second highest variable loaded on this factor is a graph. The apparent reasons for this variable's high loading are that the table requires perusal in order to appreciate its significance and that the items referring to it require a logical sort of extrapolation.

Variable 11, on the other hand, the data interpretation variable loaded lowest on the Data Interpretation factor, is an ordinary graph with ordinate and abscissa. The significance of the graph is apparent at first glance. The only complexity in the graph is that the ordinate is a logarithmic scale, but examinee perception of this fact does not appear to be essential in order to answer the items. The items ask for that which is explicitly given in the graph. They also deal with relatively difficult quantitative concepts such as rate of change.

Discussion

The item construction "rule" mentioned in the opening paragraph of this paper is that if you want to measure a specific ability with a given item type, make that aspect of the item type that distinguishes it from other item types difficult, and make those aspects of the item type that are measured by other items relatively easy. For example, antonyms and synonyms are good measures of vocabulary; therefore, it may be inefficient to use difficult vocabulary items in reading passages since this would tend to reduce the amount

of Reading Comprehension variance and would be a cumbersome way of measuring Vocabulary.

The Reading Comprehension factor in this study seems to measure the ability to appreciate the connotations and implications of what is read or to ascertain the meaning of a word from its use in context, rather than the ability to understand what is explicitly stated in the reading passage.

Analogous to appreciating connotations of reading materials, the Data Interpretation factor seems to be best identified as ability to extrapolate or draw logical inferences based upon the information presented in the graph or the table.

Since all the items included in the Selective Service test were carefully pre-tested, all the variables load at least on the General factor; however, since the General factor is well measured, variables loaded only on the General factor have little to offer for increasing a composite of the items types with the criterion of academic success. The item types used in this test are generally found to make independent contributions to college success. Since this is the case it seems we would more generally come closer to the factor structure of the criterion by avoiding items which measure only general intelligence.

As a major conclusion to be drawn from this study, the investigators would suggest that failure of one of the reading subtests to load on its group factor and failure of one of the data interpretation subtests to load on its group factor indicate that the variance explained by these two group factors is due to the intrinsic difficulty of the reading passage or graph rather than to the difficulty of the items referring to the reading passage or graph. If the items are complex and the chart or reading passage is straightforward, only the General factor will be measured, but if the graph or reading passage is difficult or complex, the appropriate group factor will be measured as well.

REFERENCES

- Thurstone, L. L. *Multiple-Factor Analysis*. Chicago: University of Chicago Press, 1947.
- Wherry, R. J. "A New Iterative Method for Correcting Erroneous Community Estimates in Factor Analysis." *Psychometrika*, XXIV (1949), 231-241.

ITEM DIFFICULTY AS A FUNCTION OF PERCEIVED ITEM DIRECTIONS^{1,2}

ROBERT PERLOFF

Purdue University

AND

LEROY WOLINS

Iowa State University

THIS study is based upon data collected in connection with Form 4 of the Selective Service College Qualification Test. Draft-registered college men are eligible to take this test; test performance is one of the criteria established by the Selective Service System in considering which Selective Service registrants should be deferred from military service until they have had an opportunity to complete one or more integral units of their advanced education.

The test consists of 150 five-alternative items, equally divided among verbal reasoning, arithmetic reasoning, reading comprehension, and data interpretation items. The data comprising this paper are derived entirely from the 38 verbal reasoning items, equally divided among synonyms, antonyms, verbal analogies, and sentence completion items.

For each of the nine synonyms or "sames" items, the respondent is required to select the word which has a meaning most nearly the same as the meaning of the stem word; for the ten antonyms or "opposites" items, the word most nearly the opposite of the meaning of the stem; for the ten analogy items, the subject must choose from among the five alternatives that pair of words whose members bear

¹ Adapted from a paper presented by the investigators at the 64th Annual Convention of the American Psychological Association, Chicago, September 5, 1956.

² Opinions expressed herein are those of the authors and do not necessarily reflect official views or policies of the Selective Service System.

a relationship to each other similar to the relationship between the two words in the stem; and for each of the nine sentence completion items the examinee is confronted with a sentence in which one word has been omitted, his task being to select one word which agrees the best with the meaning of the sentence as a whole.

It was noted both subjectively while examining a preliminary form of the verbal reasoning subtest, and empirically by studying some early item analysis data, that with respect to the "sames" and "opposites" items the subjects may have been responding to "opposites" as if, instead, the directions were to respond to "sames." While disturbing, this phenomenon did not seem to be unreasonable, since (1) "sames" and "opposites" are conceptually identical save for the instruction that for "sames" the examinee is to select a word whose meaning is most nearly the *same* as the stem word, while for "opposites" he is to select a word whose meaning is most nearly the *opposite*; (2) the "sames" item block preceded the "opposites" item block; and (3) a degree of negative transfer or interference may have been operating as a result of stimulus generalization; that is, where interference or proactive inhibition (the effect of task A upon the learning of task B) is manifest in terms of a performance decrement in the second task ("opposites") being determined by the set established by and reinforced within the first task ("sames"). Having observed this curious though hopefully explicable phenomenon, the investigators sought to examine this positioning or set effect more systematically.

Method

The verbal reasoning items were administered to three samples of subjects. The first consisted of 360 college freshman women in a southern university. This group of women was administered the verbal reasoning items for item analysis purposes and it was from this larger pool of items that the 38 verbal items for Form 4 were ultimately selected. The four verbal items-types, i.e., "sames," "opposites," analogies, and sentence completions, were scrambled in order of presentation within the test, though each of the four types of items appeared within its own item-block. In three of the four forms thus scrambled, "sames" preceded "opposites," though not necessarily *immediately* before "opposites." (The development of these three experimental forms as well as their inclusion of the four item-types in scrambled fashion, preceded the decision to conduct

this little study. Hence there are no definitive data requisite for objective comparisons of "sames" preceding "opposites" and "opposites" preceding "sames.") This college freshman women sample is designated "A" in Table 1.

The second sample was drawn randomly from the national population of selective service registrants that actually took Form 4 of the Selection Service College Qualification Test operationally in November 1955. For this group of 926 college men, representing roughly a 10 percent sample, again the "opposites" preceded the "sames." This sample is labeled "B" in Table 1.

The third sample consisted of 238 male seniors in a midwestern high school. These subjects were administered the verbal reasoning and the other three subtests, with Form 1c of the Army General Classification Test, in order to establish a partial conversion table of Selective Service test scores in terms of the AGCT score scale, a conversion required by the Selective Service System. For this sample, designated "C" in Table 1 the ten "opposites" preceded the nine "sames."

In the operational form of the test "opposites" appeared *before* "sames," since it was suspected that it would be easier to establish a "sames" set than an "opposites" set, i.e., the "sames" items being intrinsically easier to comprehend (or at least so thought the investigators) would be less susceptible to the interfering influence of a preceding task.

Unfortunately, problems of uniformity and of maximizing test administration simplicity and efficiency precluded the use of a counterbalanced design, for it would have been desirable, certainly, to have as many students faced with "sames" before "opposites" as with "opposites" before "sames." While on this and on other accounts the results to follow may be lacking in the definitiveness or rigor necessary for the establishment of lasting psychometric doctrine, the investigators believe that the results may be heuristic.

Results

The results of the study are summarized in Table 1. These average item difficulties (\bar{p}) show clearly that while "sames" may have been intrinsically more difficult than "opposites" regardless of position in the subtest, they were significantly more difficult when preceded by "opposites" (samples "B" and "C") than when followed by "opposites" (sample "A").

TABLE 1

Mean Item Difficulties of Verbal Reasoning Items Administered to High School Seniors and to College Students

Item type (number of items)	"Sames" before "Opposites" ¹		"Opposites" before "Sames" ¹			
	Sample A (<i>n</i> = 360) <i>College Freshman Women</i> ²		Sample B (<i>n</i> = 926) <i>College Men</i>		Sample C (<i>n</i> = 238) <i>High School Senior Men</i>	
	<i>p</i>	<i>SE_p</i>	<i>p</i>	<i>SE_p</i>	<i>p</i>	<i>SE_p</i>
I. "Sames" (9)	.53	.026	.48	.016	.35	.031
II. "Opposites" (10)	.59	.026	.74	.014	.61	.032
III. (II - I)	.06*	—	.26**	—	.26**	—
IV. Analogies (10)	.53	.026	.75	.014	.65	.031
V. Completions (9)	.46	.026	.60	.016	.48	.032
VI. Mean Item Difficulty (for all items except the 9 "sames")	.53	.026	.70	.015	.58	.032
			(.70 - .48)***		(.58 - .35)***	

¹ For Samples B and C, the order of item blocks was "opposites," "sames," analogies, and completions. For Sample A, the order was scrambled, although in three of the four forms "sames" preceded "opposites."

² Data on the college freshman women were gathered and analyzed by Dr. and Mrs. E. E. Cureton, Knoxville, Tennessee.

* Differences between (II, IV, and V—combined) and I.

* .06 fails to be significantly different than .00 at the 10 percent level.

** Significantly different than .00 beyond the .5 percent level.

In sample "A," three out of four subjects received forms of the verbal reasoning subtest in which "sames" preceded "opposites." The average difficulty for "opposites" was .59; for "sames," .53; yielding a difference of .06 which is different than .00 only at the 11 percent level of significance. On the one hand, therefore, the statistical evidence does not support the hypothesis that for the 360 college freshman women (sample "A") "opposites" were less difficult than "sames"; while on the other hand the evidence did not indicate that for this sample of college women the positioning effect influenced negatively mean performance on the second task ("opposites"). The validity of these results may be attenuated, however, by three sources of possible contamination, which must be considered before discarding the positioning effect hypothesis: (1) "sames" preceded "opposites" in three out of four cases, *not* in four out of four. This means that for roughly 90 or 25 percent of the 360 women in sample "A," "opposites" preceded "sames" and that this may have reduced spuriously the mean difficulty for "sames"; (2) for sample "A" there are no data to show the effect of position or set when "sames"

precede "opposites" and when "opposites" precede "sames"; and (3) there are no definitive "sames" and "opposites" item difficulty parameters, without which one cannot say with great assurance that "sames" are not indeed more intrinsically difficult than are "opposites," notwithstanding the apparent negation of this possibility as evidenced by the insignificant difference of .06.

The observed differences of .26 for both samples "B" and "C" were significantly greater than .00 beyond the .5 percent level. That is, mean "sames" performance is far beneath that of mean "opposites" performance, remembering that "opposites" preceded and therefore may have exerted a negative influence upon "sames" in sample "B" and "C." Even if it be conceded that "sames" may be inherently more difficult, they are even *more* difficult in "B" and "C" where "sames" follow opposites than in "A" where "sames" precede "opposites" in 75 percent of the cases.

In order to investigate the comparative difficulty of "sames" and that of the other verbal reasoning items combined, not including "sames," the data in the last row of Table 1 were obtained. In sample "A" "sames" appeared to be identically difficult ($p = .53$) as the other verbal reasoning items; while for samples "B" and "C," "sames" were evidently far more difficult than were the other verbal reasoning items—far more difficult, the investigators suggest, not necessarily because of inherent differences factorially between "sames" and "opposites," but rather because here, in samples "B" and "C," "sames" followed "opposites" in the verbal reasoning subtest, and the carryover of set or the presence of negative transfer or of proactive inhibition may have produced this meaningful decrement in "sames" performance.

Examining Table 1 as a whole, the investigators believe that the trends indicated encourage generally the positioning effect hypothesis, notwithstanding differences in mean item difficulties that may be attributable to well-known item shrinkages and to a certain degree of amorphousness in these data, deriving from an imperfect—though operationally unavoidable—experimental design.

Discussion

In part the investigators are encouraged by data reported by Wolins and Perloff (1956). Data collected and analyzed in connection with that particular study reveal a validity against high school grades of .45 for "opposites" and of .23 for "sames," for the

sample of 238 male high school seniors who were exposed to the "opposites" before the "sames" items. It is unreasonable to view this difference as less criterion overlap for "sames" than for "opposites." It may be reasonable, however, to suggest that aside from possibly lower reliability of the "sames" items, the reason "sames" is less valid than "opposites" is that the unintentionally measured effects of proactive inhibition or of positioning are *invalid* for high school grades, that these phenomena are neither directly nor indirectly considered when school grades are assigned. Continuing on this line of evidence from Wolins and Perloff, "opposites" load .41 while "sames" load .29 on the "General Intelligence" factor extracted in their factor analysis. Finally, it would be expected rationally that the uncorrected correlation between "sames" and "opposites" would be higher, were it not for this accidental positioning phenomenon, than the r of .36 actually obtained.

From a test construction point of view, this study suggests that in pencil-paper tests the measurement of special abilities may be contaminated by extraneous though possibly important factors (such as attention to instructions, alertness, set, carelessness, "too much" or too little motivation) unless item blocks designed to measure these abilities are adequately separated in the test or subtest and otherwise distinguished one from the other. To further reduce this postulated perseveration effect, instructions should be made more explicit and differential, both in terms of type-face differences and special textual emphases and admonitions.

Upon further verification through counterbalanced designs impractical in the current study and through the use of specially constructed items containing both synonyms and antonyms among the alternatives so that definitive alternative analyses can be conducted in order to see whether examinees indeed choose "opposites" alternatives in "sames" item blocks following "opposites" item blocks, and vice versa, the investigators suggest that this positioning phenomenon *per se* be pursued deliberately both as a device for measuring set or the ability to change one's direction and as an instrument in the study of human learning in general and of proactive inhibition in particular.

Summary

In this study an effort was made to determine the effect of a set of incorrectly perceived directions upon responses to similar verbal

reasoning items in adjacent item blocks. A possible positioning effect was observed when a college aptitude subtest of verbal reasoning was administered to a sample of college freshman women. Approximately 75 percent of the subjects were confronted with the synonyms item block before the opposites item block. An examination of item responses suggested the possibility that many of the subjects were answering the opposites items retaining erroneously the directions from the preceding item block, synonyms. In subsequent administrations of the subtest, to a sample of college men selected randomly from a national sample and to a sample of high school senior males, the opposites item block preceded the synonyms item block, operational constraints precluding a counterbalanced design for these two samples.

The results indicated that average item difficulties show clearly that while synonyms were intrinsically more difficult than the opposites regardless of position in the subtest, they were significantly more difficult when preceded by opposites than when followed by opposites.

Reasons for the positioning effect were suggested, along with an indication of possible ways in which the effect can be reduced or eliminated.

REFERENCES

- Wolins, L. and Perloff, R. "The Factorial Composition of AGCT 'Subtests,' along with College Aptitude Items and High School Grades." *American Psychologist*, XI (1956), 451. (Abstract)

EFFECTS OF THREE SETS OF TEST INSTRUCTIONS ON SCORES ON AN INTELLIGENCE SCALE¹

KAORU YAMAMOTO

AND

HENRY F. DIZNEY

Kent State University

IN the past half century, various aspects of test administration and scoring procedures have been studied since their standardization has been regarded as the cornerstone of valid and reliable testing. Effects of variation in test instructions and phrasing of questions on test results are currently receiving much attention under such topics as social desirability and response sets (Christie and Lindauer, 1963; Couch and Keniston, 1961; Cronbach, 1946; Edwards, 1957; Hills, 1961). Most such studies have been concerned with personality inventories while little or no interest has been shown to the "now-established" aptitude and achievement tests.

If, however, a testee is, in fact, influenced by general social considerations (social desirability) and by personal tendency to agree or disagree (response sets) when he reads and reacts to test questions, then he should also be thus influenced when he receives and reacts to test instructions. Therefore, if a test is introduced differently to two groups of subjects as an instrument measuring traits A and B respectively, and if trait A is socially more desirable than trait B, it should then be predicted that subjects will be differentially motivated, resulting in differential achievement, under the two instructional conditions. Such results, moreover, should not be limited to tests of personality alone.

¹ Cooperation of the Mentor, Ohio, School District (Superintendent, Dr. E. Masonbrink) in this study is gratefully acknowledged.

In the present study, an intelligence scale was administered to elementary and high school pupils under three sets of instructions. In one group, the test was introduced as a test of intelligence, while, in another group, the same test was introduced as a test of achievement. In the third group, the test was not labeled at all and introduced as a "routine" test. According to the reasoning stated above and in view of the high social prestige and premium given to a "high IQ," it was predicted that the subjects would gain the highest scores under the "intelligence" instruction, followed by scores under the "achievement" and "routine" instructions in that order. Since the differential prestige is undoubtedly a product of social learning, it was further predicted that the differences in test scores would increase with age of subjects and that girls would produce larger score differences than boys. It is now well known that social conformity and dependence are emphasized much more in raising girls than in raising boys in our society.

Procedures

Subjects

A total of 557 children from the fourth, seventh, tenth, and twelfth grades of a large urban school system in northeastern Ohio participated in the study. In each grade, subjects were randomly assigned to one of the three groups called respectively "Intelligence," "Achievement," and "Routine." The three groups in each grade were then administered the same test under different instructions in three different rooms by the authors and their graduate assistants.

Instrument

The Kuhlmann-Anderson Test, Seventh Edition, was chosen as the basic instrument. This edition does not carry the word "intelligence" in its title and no mention of intelligence or IQ is made in its preliminary instructions or directions for subtests. The contents of the test could as well be described by "a measure of intelligence" as by "a measure of achievement." Booklets (form) D, G, H, were used for the fourth, seventh, tenth, and twelfth graders, respectively. The results were expressed as deviation IQ's based upon a mean of 100 and a standard deviation of 16 (Anderson, 1960).

Instructions

The first paragraph of the "Preliminary Instructions to the Students" of the Kuhlmann-Anderson Test² was replaced by one of the following three paragraphs. In addition, subjects were reminded briefly of the nature of their test (intelligence, achievement, or routine) at the end of the preliminary instructions³ and again at the conclusion of the fourth subtest⁴ (there are eight subtests in each form). Except for these three places, no alteration was made in the original instructions.

Intelligence. "The test you are going to take today is a test of *intelligence* and a very good one, too. As you know, a test of *intelligence* will give us your *IQ* which tells us how *smart* you are, how *clever* you are, and how *capable* you are. Your *IQ* also tells us how *bright* you are compared to other students (pupils) like yourself. A *high IQ* is necessary to make good grades in school, to learn things quickly and well, to go to college, and to hold a good job. A *bright, able* student (pupil) makes a high score on this test. Listen carefully to the directions I am going to give you and do your best to get a *high IQ, O.K.?*"

Achievement. "The test you are going to take today is a test of *achievement* and a very good one, too. As you know, a test of *achievement* will tell us how *well* you have *learned* your lessons, how much you *know* about many things, and how *skillfully* you can use what you know. Your *achievement* also tells us how *much* you have *learned* compared to other students (pupils) like yourself. A *good, hard-working* student (pupil) makes a *high score* on this test. Listen carefully to the directions I am going to give you and do your best to get a *high achievement score, O.K.?*"

Routine. "This is just a *routine testing* on how well you under-

² The paragraph, in original, reads as follows (booklets G and H): "This is a test that will sample your information and problem solving skill of different kinds. Your scores will be helpful to your teachers in planning your work and to you and to your advisors in making decisions about your courses and future plans. Your record will be determined in large part by how carefully you listen to directions and how carefully you follow them. Be sure to listen attentively to the directions, for they can be given only once." A simpler wording is used in booklet D.

³ For example, "Remember, this is a test of your *IQ*. Do your best and get a *high IQ*."

⁴ For example, "Are you doing your best? You must do your best to get a *high IQ* on this test, O.K.?"

stand words, their meanings and uses, and how well you can answer questions and do work with numbers and figures. Listen carefully to the directions I am going to give you and do your best to get a high score, O.K.?"

Results

Mean IQ's obtained by three groups of subjects are shown in Table 1.

When the IQ's presented in Table 1 were analyzed by an analysis of variance (method of unweighted means), the results shown in Table 2 were obtained.

From Table 2, it is observed that the effects of instructional set and of sex are significant, while grade effects and four interaction effects do not reach the chosen level of statistical significance ($\alpha = .05$). Scheffé's test, applied to the three instructional set means, revealed that the mean obtained under the Intelligence in-

TABLE 1
Mean IQ on Kuhlmann-Anderson Test by Instructional Set, Grade, and Sex

Grade	Instructional Set	Sex	N	Mean IQ
4	Intelligence	Boy	28	112.2
		Girl	20	121.1
	Achievement	Boy	27	107.7
		Girl	17	115.5
	Routine	Boy	21	104.8
		Girl	20	109.5
7	Intelligence	Boy	36	118.9
		Girl	22	116.9
	Achievement	Boy	30	108.5
		Girl	27	113.4
	Routine	Boy	15	109.3
		Girl	20	114.7
10	Intelligence	Boy	30	114.0
		Girl	29	112.1
	Achievement	Boy	18	115.9
		Girl	31	111.9
	Routine	Boy	21	105.0
		Girl	15	107.9
12	Intelligence	Boy	28	115.4
		Girl	23	121.7
	Achievement	Boy	27	112.9
		Girl	16	115.8
	Routine	Boy	15	109.8
		Girl	21	112.3
Total			557	

TABLE 2

Analysis of Variance (Method of Unweighted Means) of Mean IQ on Kuhlmann-Anderson Test

Source of Variation	df	MS	F	p
Instructional Set (i)	2	106.64	14.8	<.01
Grade (g)	3	15.76	2.2	NS*
Sex (s)	1	61.44	8.5	<.01
(i) × (g)	6	9.37	1.3	NS
(i) × (s)	2	2.89	1>	NS
(g) × (s)	3	16.65	2.3	NS*
(i) × (g) × (s)	6	5.61	1>	NS
Residual	533	7.21		

* .05 < p < .10.

struction (116.3; $N = 216$) is significantly larger ($p < .01$) than the Achievement mean (112.1; $N = 193$), while the latter is not significantly different from the Routine mean (109.1; $N = 148$). Between two sexes, girls obtained a higher mean IQ (114.4; $N = 261$) than that of boys (111.7; $N = 296$).

Discussion

It is noted from the above results that the authors' original prediction was *partially* confirmed. Subjects did in fact react differentially to different instructional sets, although the predicted longitudinal and sex differences were not observed (no interactions significant). In this particular group of subjects, girls, on the average, performed significantly better than boys regardless of the specific instructions given.

It is further seen that the significant instructional set effects stem from the significant difference in IQ between the Intelligence group and the Achievement-Routine groups. Seemingly, then, the concept of IQ, whatever it means to these students, enjoys higher social prestige than the concept of achievement and thus tends to motivate subjects to better test performance. It is rather surprising to find no difference between the Achievement group and the Routine group. Either the concept of high achievement did not hold any special meaning and implications for these students or this concept was subjugated to a more inclusive concept of a "higher score" on a test. In this regard, it is regretted that a group was not formed to be administered the test without any alterations at all in the original instructions.

Pending replications, the study would seem to suggest many interesting topics for further research. It was, for example, observed that the use of terms such as "intelligence," "IQ," "bright," and "able" in test instructions makes a difference in subjects' performance. Now, there are many aptitude tests which explicitly employ these terms in their titles and instructions, while there are some which do not mention anything of the sort. Are these two kinds of tests then actually measuring the same thing or are they measuring different qualities? Do such subtle, social pressures operate in other kinds of psychological measurements than the "objective" tests of intelligence?

Why did the prediction of larger differential effects for older subjects and for girls fail? Was this because the social desirability of high intelligence had been already well established among both sexes by fourth grade? Do we then find the expected results in subjects younger than the present ones? Or was this because the two sexes in the subject group were unbalanced in terms of their mental aptitudes as shown in the significant sex effects? Would we have found the expected results if we had matched the basic potentials in boys and girls?

Since the observed difference in test performance is ascribed to the general social reinforcement factor attached to "IQ," it will be interesting to study the possible differences among subjects from different social classes and residential areas (urban, rural, etc.) in this and other societies. It is, for example, readily hypothesized that the observed differential effects of test instruction would be larger in middle-class students than in lower-class students because of the well known middle-class value system which assigns a premium to "smartness" and the resultant upward social mobility.

In a similar vein, it will be informative to study whether the observed differential effects are found along all ranges of mental aptitude. So called "gifted" children could be studied by comparing, for example, a group of high IQ children who *know* that they have high intelligence with another group of high IQ children who *do not know*. Here it could be predicted that the latter group will show larger instructional set effects because, for them, a high IQ is still an aspired quality, while for the former group this is an established fact. As another example, a group of mentally-retarded children could be compared with a group of MA-matched normal children. Because of the reported observation (Stevenson, 1962) that re-

tardates are more affected by social reinforcements than normals, it is to be predicted that the differential effects will be larger among retarded children than among normal children.

Concepts other than "intelligence" and "achievement" could be studied for their social acceptance and prestige. For example, the now popular concept of "creativity" might be studied in relation to the old guard, "intelligence," among various groups of subjects. Possible variations are limitless.

Finally, individual differences should be carefully studied. Who is affected by social factors in test performance and who is not? And, of course, why the difference?

Summary

The Kuhlmann-Anderson Test was administered to a total of 557 subjects of fourth, seventh, tenth, and twelfth grades under three different instructional sets. In each grade, subjects were randomly assigned to three groups and the test was introduced to these groups as a test of *intelligence*, a test of *achievement*, and a *routine* test, respectively. Because of the differential social emphases upon intelligence and achievement, it was predicted that (1) performance of the first group (Intelligence) would be the best, followed by the second group (Achievement), and the third (Routine), in said order; (2) that the preceding differences would be greater as subjects became older; and (3) that sex differences would interact with grade level.

When the results were converted to deviation IQ's and analyzed by an analysis of variance, it was found that (1) the mean IQ obtained by the Intelligence group (116.3) was significantly higher than those obtained by the Achievement group (112.1) and the Routine group (109.1), while the latter two group means were not significantly different; (2) girls obtained a mean IQ (114.4) significantly higher than that (111.7) obtained by boys; and (3) no other effects, including the (instructional set-grade) interaction and the (instructional set-sex) interaction, were statistically significant.

Possibilities for further studies were discussed and various approaches were proposed.

REFERENCES

- Anderson, R. G. *Technical Manual for Kuhlmann-Anderson Test, Seventh Edition, Booklets G and H (also D)*. Princeton: Personnel Press, Inc., 1960.

- Christie, R. and Lindauer, Florence. "Personality Structure." In P. R. Farnsworth, Olga McNemar, and Q. McNemar (Editors), *Annual Review of Psychology* (Volume XIV). Palo Alto, California: Annual Reviews, 1963, 201-230.
- Couch, A. and Keniston, K. "Agreeing Response Set and Social Desirability." *Journal of Abnormal and Social Psychology*, LXII (1961), 151-174.
- Cronbach, L. J. "Response Sets and Test Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 475-494.
- Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden Press, 1957.
- Hills, J. R. "The Influence of Instructions on Personality Scores." *Journal of Counseling Psychology*, VIII (1961), 43-48.
- Stevenson, H. W. "Discrimination Learning." In N. R. Ellis (Editor), *Handbook of Research in Mental Retardation*. New York: McGraw-Hill Book Company, 1962.

THE RELATIONSHIP BETWEEN PERSISTENCE, INSOLENCE, AND PERFORMANCE, AS A FUNCTION OF GENERAL ABILITY

DAVID KIPNIS¹

Navy Medical Research Institute,
Bethesda, Md.

THE present lack of stability in the relationship between personality measures and behavior has made it difficult to use personality tests for predictive and counseling purposes in the same way that intelligence tests are used. For instance, Spielberger (in press) has noted that investigations of the relationships between personality measures and academic performance have resulted in equivocal and inconsistent findings. In some instances a given personality measure has been found to be positively related to school grades and in other instances the same measure has shown negative relationships. Thus, a continuing problem in the measurement of personality has been the identification of conditions under which a given personality trait is related to some specified behavior.

Considering all the possible situational and intra-individual factors that may affect relationships between personality and performance, the problem is of course a monumental one. On the other hand, beginnings of an attack on the problem are being made in several areas. Thus, several investigations have been concerned with the differing relationships between personality and performance as a function of differences in leadership style (Fiedler, 1962; Haythorn, et al., 1956), of differences in the social setting (McGrath, 1962) or of differences in the property of tasks subjects must perform (Feather, 1963; Taylor and Spence, 1952). Other investigators (e.g., du Mas, 1958; McQuitty, 1957) have been concerned with the question of whether interactions between traits within

¹ The opinions expressed here do not necessarily represent those of the U. S. Navy.

an individual may modify the relationship between a given test variable and subsequent performance on criterion tasks. Most recently the formal introduction of the concept of moderator variable by Saunders (1956) has stimulated considerable interest in the detection of trait interactions (Berdie, 1961; Fredericksen and Melville, 1954; Ghiselli, 1956; Kipnis, 1962). Ghiselli (1963) has recently discussed the theoretical issues involved in the use of moderator variables.

The moderating effects of general intelligence on the relationship between persistence and performance have been reported by Kipnis (1962), and serve as a basis for the research to be presented here. In the previous study it was found that a measure of persistence beyond minimum standards called the Hand Skills Test, predicted the school and job performance of Navy enlisted men who were below the median in general intelligence for their Navy occupational grouping. Persistence was not related to performance among men who were above the median in intelligence. Subsequent unpublished analysis of a second test called the Insolence Scale (initially discussed in Kipnis and Glickman, 1962), administered at the same time as the Hand Skills Test, found it predicted performance among men who were above the median in general intelligence for their occupational groupings, but did not predict among men who were below the median in general intelligence. This report presents results of an attempt to extend these findings to larger samples. Two hypotheses were tested.

1. The relationship between persistence, as measured by the Hand Skills Test, and school and job performance is a function of general intelligence. Among lower ability men persistence will be positively related to performance. Among higher ability men persistence will not be related to performance.
2. The relationship between the Insolence Scale and school and job performance is a function of general intelligence. Among lower ability men, the Insolence Scale will not be related to performance. Among higher ability men, the Insolence Scale will be negatively related to performance.

Procedure

Samples

The general design of the study was to test Navy recruits two weeks before their entrance into one of six Navy trade schools. Ap-

proximately one and a half years later evaluations of their performance were obtained from their supervisors. At the time of testing Ss had been in the Navy nine weeks and over 90 percent ranged in age from 17 to 18 years. Ss entering the following Navy trade schools were used.

1. Electricians Mate (*EM*). Three hundred and sixty-nine men were initially tested. Electrician Mate's school grades were obtained for 331 *EMs* and job performance evaluations were obtained for 313 *EMs*. The *EMs* had been on the job for an average of nine months when their performances were evaluated.

2. Radiomen (*RM*). Three hundred and seventy-six *RMs* were initially tested. Radiomen school grades were obtained for 372 *RMs* and job performance evaluations for 281. The *RMs* had been on the job for an average of seven months at the time their performances were evaluated.

3. Fire Control Technicians (*FT*). Four hundred and one *FTs* were initially tested. Grades at Fire Control school were obtained for 364 *FTs* and job performance evaluations were obtained for 335 *FTs*. Average length of time on the job was six months.

4. Hospital Corpsman (*HM*). Three hundred and sixty *HMs* were tested; Hospital Corpsman school grades were obtained for 358 *HMs* and job performance evaluations for 313 *HMs*. Average time on the job was seven months.

5. Interior Communications Specialists (*IC*). Three hundred and sixty-nine *ICs* were tested; school grades for 367 *ICs*, and job performance evaluations for 289 *ICs* were obtained. Average time on the job was eight months.

6. Machinist Mates (*MM*). Two hundred and eighty-two *MMs* were tested; school grades for 267 *MMs*, and job performance evaluations for 241 *MMs* were obtained. Average time on the job was nine months.

Tests and Criterion Measures

1. Hand Skills Test. The test sought to measure motivation to persist beyond minimum standards on tiring tasks. It consists of sequentially numbered boxes in which examinees pencil tally marks (■). The test rapidly promotes hand and arm fatigue and is presented to Ss as a measure of how rapidly people can use their hands and fingers. It has a one minute practice session and three parts of four minutes each. A "passing score" is announced prior to each of

the four minute parts.² Pretesting had established that this score could be reached by all examinees in the time allowed.

The test seeks to discriminate between those who stop and slow down after the passing score is reached and those who continue to strive. The score used is: number completed in part three minus number completed in the practice session.

2. Insolence Scale. This scale is derived from a modification of the Risk Scale developed by Torrance and Ziller (1957). In an earlier study, contrary to expectations, 27 of the 58 items from the Risk Scale had been found to be negatively related to the job performance of enlisted men. These items, labeled the Insolence Scale, subsequently showed higher validity than the total Risk Scale with a new sample of enlisted men (Kipnis and Glickman, 1962). In general, the 27 items portray physically active, aggressive, somewhat hostile and reckless personalities, who early in life had become independent of family and school control, and who it is suspected continued to maintain an independent and rebellious attitude toward most attempts at controlling their behavior. In terms of character structure, the items appear most descriptive of individuals who would be diagnosed as passive-aggressive personalities.

The test is negatively keyed so that high scores denote men high in insolence.

3. General Classification Test (GCT). The GCT was used as a measure of general intelligence. The GCT is the measure of verbal reasoning ability in the Navy's Basic Test Battery and is used to classify enlisted men when they first enter the Navy.

4. School grade. At each trade school, grades were expressed on a scale ranging from 60 to 100. Ss flunking from school for academic or motivational reasons were assigned arbitrary scores of 59. Within each school, two criterion groups were formed, based upon the distribution of school grades—a below average school group (men whose school grades fell in approximately the bottom third of the distribution) and an average school group, composed of all other men in the school.

² Instructions for the first four minute part were "You have four minutes to complete part two. Fill as many boxes as you can. Remember each box must be completed before the next one is started. The more you do the better your score. In order to pass this part of the test you must make at least 100 boxes. Less than 100 is a failing score." Subsequent parts of the test raised passing scores by five, and limited instructions to only announcing a new passing score and the time allowed.

5. Job performance evaluations. Sixteen to eighteen months after testing, supervisors were requested to evaluate each *S* on four 14-point scales covering the following areas: (1) Technical Competence; (2) Willingness to Work; (3) Respect for Authority; (4) Overall Evaluation of the man's worth to the Navy. Within each sample, and for each performance area in turn, two groups were formed consisting of men categorized as below average (approximately the bottom third of each distribution of evaluations) and men categorized as average or better.

Analysis

The data were organized for analysis in the following way.

1. Based upon the distribution of GCT scores, *Ss* within each sample were divided into quartiles. The *EMs*, *HMs*, *ICs*, *MMs*, and *RM*s had approximately the same distributions of GCT scores, and consequently the cutting points for GCT quartile for these five samples were within one point of each other. On the other hand, the mean GCT score for *FTs* was approximately one standard deviation higher than the GCT means for the other five samples. As a result, many of the *FTs* who were assigned to the lowest GCT quartile for *FTs*, had in fact GCT scores as high as men in the other five samples who were assigned to the 50th-74th GCT quartile. The elevated GCT scores of *FTs* are due to the fact that their work and training places heavy intellectual demands upon the individual, and consequently the Navy selects *FTs* from among its brightest recruits.

2. Within each sample, biserial correlations were computed between the experimental tests and the school and job criteria. Correlations were computed at each ability level and for each total sample, ignoring ability level.

Restriction in Criterion Variance

For all six samples, statistically significant correlations ranging from .24 to .40 were obtained between GCT scores and school grades. As a result it was not possible to categorize equal numbers of men in the highest and lowest GCT quartiles as being below average in school grades. The number of *Ss* in each sample classified as below average in school grades in the highest GCT quartiles ranged from 12 percent to 27 percent (median 18 percent). The comparable fig-

ure for *Ss* in the lowest GCT quartile classified as below average in school grades ranged from 44 to 61 percent (median 51 percent).

There were no statistically significant correlations between GCT scores and job performance evaluations. Consequently, relatively equal numbers of men in each GCT quartile could be categorized as below average on the four job performance evaluation scales.

Restriction in Experimental Test Variances

For each sample the product-moment correlations between the two experimental tests and GCT scores by ability level, and for each total sample were nil. Table 1 gives these correlations. Since the experimental tests were not correlated with GCT there was no systematic restriction in the range of test scores when *Ss* were subdivided into GCT quartiles.

Results

The results for school performance are given in Table 2 which shows the biserial correlations between the Hand Skills Test, the Insolence Scale, and school grades for each ability level.

Of the 24 validity coefficients for the Hand Skills Test, three were significant beyond the .01 level and one beyond the .05 level. Inspection of the distribution of these significant validities indicates slight support for the first hypothesis. The Hand Skills Test was significantly correlated with school grades among *MMs* and *HMs* who were in the lowest GCT quartiles (r 's = .43 and .36 respectively). While not predicted but consistent with Hypothesis 1, a significant negative correlation between the Hand Skills Test and school grades was obtained for *RM*s in the highest GCT quartile ($r = -.38$). Results opposite to those predicted however, were obtained in the *EM* sample. For this group, a significant positive correlation ($r = .41, p < .01$) was obtained between the Hand Skills Test and school grades for *EM*s in the highest GCT quartile.

With reference to the Insolence Scale, only three of the 24 correlations with school grades were statistically reliable beyond the .05 level. Consistent with the second hypothesis, none of the three correlations were obtained among men in the lowest GCT quartile. However, since this number of significant correlations could easily have arisen through chance, it is concluded that the data do not support Hypothesis 2.

TABLE 1
Product-Moment Correlations of GCT with the Hand Skills Test (Hand) and the Insolence Scale (Insol)
(By GCT Quartile)

GCT Quartile	IC		EM		MM		RM		HM		FT	
	Insol	Hand	Insol	Hand	Insol	Hand	Insol	Hand	Insol	Hand	Insol	Hand
1-24	.00	-.12	.01	.06	-.01	.10	-.03	.15	.08	.01	.07	.13
25-49	-.01	.04	.18	.09	.07	-.05	-.01	-.16	.05	.15	.01	.03
50-74	-.01	.13	-.11	.14	-.01	.10	.03	.11	.02	-.05	.09	-.01
75-100	-.07	-.18	-.11	.04	.10	.02	-.03	-.08	-.15	.04	.00	-.03
Total Sample	.04	.07	.01	.04	.08	.06	-.01	.08	.03	.04	.04	.14

TABLE 2
Biserial Correlations of School Grades with the Hand Skills Test (Hand), and the Insolence Scale (Insol)
(By GCT Quartiles)

GCT Quartile	IC		EM		MM		RM		HM		FT	
	Insol	Hand	Insol	Hand	Insol	Hand	Insol	Hand	Insol	Hand	Insol	Hand
1-24	-.09	-.12	-.15	-.17	-.18	.43**	-.09	-.03	-.12	.36**	-.12	.10
25-49	-.22	-.14	-.04	.01	-.05	-.04	-.26*	.01	-.04	-.21	-.18	.03
50-74	-.14	.14	-.28*	-.13	-.18	.06	-.10	-.05	-.16	.16	-.20	-.03
75-100	.00	.25	-.05	.41**	-.04	-.05	-.20	-.38*	-.53**	.08	-.02	.01

* $p < .05$.** $p < .01$.

Table 3 gives the biserial correlations between the Hand Skills Test and job performance evaluations for the six samples by ability level. The Hand Skills Test correlated .55 ($p < .01$) with evaluations of Technical Competence among *RM*s in the lowest GCT quartile. Similar findings for *RM*s were reported in the previous study (Kipnis, 1962). However, there were no significant correlations among lower ability *S*s in the other five samples. Thus with reference to the prediction of job performance, it is concluded that the data did not support Hypothesis 1.

Table 4 presents the biserial validities for the Insolence Scale. In general, the data provide support for Hypothesis 2. Of the 20 validity coefficients that were negatively related to performance beyond the .05 level, only one was found among *S*s in the lowest GCT quartile, five were found in the 26th–50th quartile, and seven significant correlations were found in each of the two highest GCT quartiles. Finally, there was a significant positive correlation of .36 ($p < .05$) between the Insolence Scale and Evaluations of Overall Worth among *MM*s in the lowest GCT quartile. This latter correla-

TABLE 3

Biserial Correlations of the Hand Skills Test with Job Performance Evaluations (By GCT Quartile)

GCT Quartile	IC	EM	Samples		HM	FT
			MM	RM		
Evaluation of Technical Competence						
1-24	-.07	-.02	-.03	.55**	.05	-.03
25-49	.15	.05	.17	.10	-.04	.17
50-74	-.21	.22	.15	-.10	.18	-.03
75-100	.08	.11	-.07	.18	-.23	-.07
Evaluation of Willingness to Work						
1-24	.16	-.17	.14	.31	-.07	-.04
25-49	.04	.09	-.10	.00	-.10	.13
50-74	-.21	-.12	.43**	-.02	.22	-.21
75-100	.09	-.03	-.10	-.08	-.17	.12
Evaluation of Respect for Authority						
1-24	-.14	-.13	-.06	.26	-.11	-.15
25-49	.16	-.10	.09	-.09	-.16	.20
50-74	-.23	.03	.26	-.10	.15	.01
75-100	.03	.07	-.01	.11	-.16	.11
Evaluation of Overall Worth						
1-24	-.01	-.09	-.02	.26	-.07	-.03
25-49	.06	.11	.01	.10	-.04	.05
50-74	-.22	.16	.17	.13	.21	.07
75-100	.15	-.02	.02	-.01	-.17	.04

** $P < .01$.

TABLE 4

*Biserial Correlations of the Insolence Scale with Job Performance
(By GCT Quartiles)*

GCT Quartile	IC	EM	Samples				FT
			MM	RM	HM		
Evaluation of Technical Competence							
1-24	.26	.04	.24	-.05	-.24		-.08
25-49	-.12	-.16	-.13	-.08	-.31*		-.06
50-74	-.06	-.21	-.07	-.07	-.24		-.13
75-100	-.01	-.20	-.30*	-.01	-.24		.03
Evaluation of Willingness to Work							
1-24	-.02	-.02	.26	-.18	-.09		-.09
25-49	-.16	-.03	.07	-.13	-.33**		-.04
50-74	.04	-.18	.11	-.24*	-.31*		-.15
75-100	.03	-.18	-.20	-.01	-.31*		-.03
Evaluation of Respect for Authority							
1-24	-.06	.01	.22	-.14	-.21		-.27*
25-49	-.08	-.09	-.06	-.07	-.46**		-.34*
50-74	-.11	-.35**	.03	-.30*	-.22		-.25*
75-100	-.09	-.12	-.33*	-.26*	-.26*		-.07
Evaluation of Overall Worth							
1-24	.15	.13	.36*	-.05	-.09		-.12
25-49	-.17	-.13	.09	.00	-.34*		-.13
50-74	.03	-.31**	-.05	-.16	-.26*		-.06
75-100	-.16	-.22	-.28*	-.22	-.33*		-.03

* $P < .05$.

** $p < .01$.

tion is consistent with Hypothesis 2, and suggests the possibility that under some circumstances high Insolence Scale scores may facilitate performance.

From Table 4 it can also be seen that the Insolence Scale was most clearly related to evaluations of Respect for Authority and evaluations of Overall Worth. Of the 21 significant validity coefficients, nine were correlated with Respect for Authority and six were correlated with evaluations of Overall Worth.

The validities in Table 4 were converted to their corresponding z 's and subjected to a Type 1 mixed analysis of variance described by Lindquist (1953). GCT quartiles were used as the first factor and performance evaluation scales as the second, with cell entries consisting of the converted correlations from each of the six samples. To eliminate negative scores, a constant of 50 was added to each converted correlation.³

Overall the analysis showed no significant relationships between GCT quartiles and Insolence Scale validity. Two further analyses

³ I wish to thank W. W. Haythorn for suggesting this analysis.

were done at this point. Analysis of variance (Type 1) was applied to the two rating scales—Respect for Authority and Overall Worth—which the Insolence Scale had predicted at greater than chance levels. This analysis yielded a significant interaction between Rating Scales and GCT ($F = 3.14$, $p < .05$, df 3 and 20), indicating that the variations in correlations between the four GCT quartiles were statistically reliable for evaluations of Overall Worth, but not for evaluations of Respect for Authority. The second analysis was based upon the observation that for all evaluations scales, the most abrupt increase in negative validity occurred between the first and second GCT quartiles. Since these trends in test validity were consistent with Hypothesis 2, a mixed type analysis of variance of the differences between correlations in the lowest GCT quartile and the remaining three GCT quartiles were computed for all rating scales. These results are summarized in Table 5 and show that the validities of the Insolence Scale in the lowest GCT quartiles were significantly poorer ($p < .05$) than the validities found in the upper three GCT quartiles.⁴

Finally, the significant F value for Evaluation Scales in Table 5 is consistent with the observation that the Insolence Scale predicted at lower levels for evaluations of Technical Competence and evaluations of Willingness to Work.

Discussion

The prior work (Kipnis, 1962) found that the Hand Skills Test predicted school grades among lower aptitude men in two out of three samples, and the job performance evaluations of lower apti-

TABLE 5

Analysis of Variance: Comparison of Insolence Scale Validities Lowest GCT Quartile vs. the Upper Three GCT Quartiles

Source	df	MS	F
Between GCT	1	413,671	6.26*
Between subjects in same group	22	66,107	
Between evaluations	3	29,385	5.13**
GCT \times evaluation	3	13,021	2.27
Interaction: pooled subjects \times evaluation	66	5,726	

* $p < .05$.

** $p < .01$.

⁴ A similar analysis comparing the validities in the lowest GCT quartile with the validities in the upper three GCT quartiles, but excluding evaluations of Overall Worth yield a significant F for GCT between the .05-.10 levels of confidence ($F = 3.40$, df 1 and 17).

tude men in four out of four samples. The present study however, provides limited support for the generality of those results. While the Hand Skills Test's prediction of school grades in the *RM*, *HM*, *MM* samples were consistent with the first hypothesis, the results among *EMs* were a reversal of the predicted relationship. Furthermore, with the exception of the results obtained among *RM*s, no support for Hypothesis 1 was obtained from the job performance analysis. On the other hand, the job performance analysis did provide some support for the second hypothesis that the Insolence Scale would be more valid among higher ability men. Seventy-one percent of the Scale's significant validities were in the two highest GCT quartiles and 29 percent were in the two lowest GCT quartiles.⁵ The findings from the analyses of variance were also consistent with Hypothesis 2, although they require replication because of the post hoc nature of the analysis. These results did suggest, however, that there was a linear relationship between intelligence and the magnitude of the validity of the Insolence Scale when an *Ss* overall worth was evaluated, and that the Insolence Scale was generally less predictive among *Ss* in the lowest GCT quartile.

One of the major problem areas of passive-aggressive personalities is their inability to accept authority. Consistent with the belief that the Insolence Scale is a measure of passive-aggressivity is the finding that the largest clustering of significant validities for this scale were with Evaluations of Respect for Authority. The fewest significant correlations were with evaluations of Technical Competence. These findings seem to be in the same direction as those of several investigators (Rasmussen, 1961; Whitman, Trosman, and Koenig, 1954) who have suggested that passive-aggressives encounter their main difficulties in their unwillingness, rather than their inability, to accept the social role which is required of them. It is fairly well agreed that if the social situation contained sources of stimulation adequate to motivate them, passive-aggressives would make reasonably satisfactory adjustments. The problem of what constitutes adequate sources of motivation for this group is, of course, the crux of the matter. Some possibilities will be examined in the context of the discussion below.

⁵ The positive correlation of .36 between the Insolence Scale and Evaluations of Overall Worth among *MMs* in the lowest GCT quartile, is considered equivalent to a significant negative correlation in the two highest GCT quartiles.

Perhaps the most perplexing question raised by the study has to do with the role of ability as a moderator variable. Several studies, both experimental (Block, 1962; Myers, Murphy, and Smith, 1963), and psychometric (Bass, et al., 1963; Goodstein and Heibrun 1962; Kipnis, 1962) have reported differing relationships between independent and dependent variables as a function of intelligence. Given the central role of cognitive processes in mediating between the individual and his environment, it may well be that many of the inconsistencies found in the relationships between personality and behavior can be clarified by a more critical examination of the role played by intelligence.

Within the context of the present study, at least two alternate explanations for the effects of intelligence seem possible. On the one hand the results may represent a genuine interaction with cognitive processes. That is, differing cognitive processes are employed by the more intelligent than the less intelligent (Kendler and Kendler, 1959; Osler and Trautman, 1961). Such differences in cognitive processes then, interact with personality to produce the differential relationships reported here. Thus, possibly insightful thinkers, who are high in insolence, lose interest rather quickly in their work; whereas individuals high in insolence, who generally rely on trial and error approaches to problem solving, retain interest in their work and concurrently project a more sympathetic facade to authority figures.

Another possible explanation is that the results represent an interaction between personality and task difficulty. Within the context of this explanation, the function of intelligence could be viewed as defining how difficult the work will be for the individual to grasp. That is, higher ability individuals will generally find a given task easier to solve than lower ability individuals. For example, a student in business administration with a modestly high IQ would have to put forth less effort to obtain a C average than would a student whose IQ barely exceeded the entrance requirements for the school. On the other hand, if the modestly high IQ student transferred to a course in theoretical physics, he might find that his IQ were somewhat below the median IQ of his new fellow students, and he might also have to exert considerable effort to maintain even a C average.

To the extent then that the work was relatively easy for higher ability men in the six samples, the possibility exists that Ss, who were also high in insolence, became rapidly bored with their work

and expressed their boredom by directing hostility toward authority. On the other hand, lower ability men probably had to keep "plugging" away to maintain an average performance level. Under these conditions, the interest of high insolent individuals may have been sustained by the challenge of work and their relationships with authorities did not deteriorate. The reverse of this argument may be applied to the Hand Skills Test results. That is, as the work became difficult for some individuals because of their relatively low ability, at times motivational variables such as persistence helped elevate their level of performance.

One of the main issues between the alternate explanations offered above concerns the stability of relationships between test scores and performance for a given individual. On the one hand, if the results represent an interaction with intelligence, the prediction is that while the magnitude or sign of the relationships will differ for persons at differing intelligence levels, test-performance relationships will remain invariant for a given individual. On the other hand, if the results represent an interaction with task difficulty, the expectation is that test scores for an individual may have differing relationship with performance, depending upon how difficult the performance requirements are for the individual to grasp.

Experimental work is being done at this time to test these alternatives. It would appear that support for the task difficulty hypothesis has some interesting implications for reducing failures attributable to the kinds of motivations considered here. For instance, one might consider placing individuals with passive-aggressive character structures in work that would be somewhat challenging or difficult for them. Assigning these individuals to easier work may lead to a failure to realize their full potential because they become "cocky" and possibly indifferent to the work assigned. In actual practice, assignment might take the form of placing these men in work that requires ability levels marginally higher than they possess.

REFERENCES

- Bass, B. M., Duntzman, G., Frye, R., Vidulich, R., and Wambach, H. "Self Interaction and Task Orientation Inventory Scores Associated with Overt Behavior and Personal Factors." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIII (1963), 101-115.
- Berdie, R. F. "Intra-Individual Variability and Predictability."

- EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXI (1961), 663-676.
- Block, C. H. "Interrelations of Stress and Anxiety in Determining Problem Solving Performance." Technical Report #8, NONR 609 (20). Yale University; Department of Psychology and Industrial Administration, 1962.
- du Mas, F. M. "Concept of the Intratest and Some Implications for Psychometric Theory." *Psychological Reports*, IV (1958), 187-192.
- Feather, N. T. "The Relationship of Expectation of Success to Reported Probability, Task Structure, and Achievement Related Motivation." *Journal of Abnormal and Social Psychology*, LXVI (1963), 231-238.
- Fredericksen, N. and Melville, S. D. "Differential Predictability in the Use of Test Scores." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XIV (1954), 647-656.
- Fiedler, F. E. "Leaders Attitude Group Climate and Group Creativity." *Journal of Abnormal and Social Psychology*, LXV (1962), 308-318.
- Ghiselli, E. E. "Differentiation of Individuals in Terms of Their Predictability." *Journal of Applied Psychology*, XL (1956), 374-377.
- Ghiselli, E. E. "Moderating Effects and Differential Reliability and Validity." *Journal of Applied Psychology*, XLVII (1963), 81-86.
- Goodstein, L. D. and Heibrun, A. B. "Prediction of College Achievement from the Edwards Personal Preference Schedule at Three Levels of Intellectual Ability." *Journal of Applied Psychology*, XLVI (1962), 317-320.
- Haythorn, W., Couch, I., Haefner, P., Langham, G., and Carter, L. F. "The Effects of Varying Combinations of Authoritarian and Equalitarian Leaders and Followers." *Journal of Applied Psychology*, LIII (1956), 210-219.
- Kendler, T. S. and Kendler, H. H. "Reversal and Nonreversal Shifts in Kindergarten Children." *Journal of Experimental Psychology*, LVIII (1959), 56-60.
- Kipnis, D. "A Noncognitive Correlate of Performance among Lower Aptitude Men." *Journal of Applied Psychology*, XLVI (1962), 76-80.
- Kipnis, D. and Glickman, A. S. "The Prediction of Job Performance." *Journal of Applied Psychology*, XLVI (1962), 50-56.
- Linquist, E. F. *Design and Analysis of Experiments*. Boston: Houghton Mifflin Co., 1953.
- McGrath, J. E. "The Influence of Positive Interpersonal Relations on Adjustment and Effectiveness in Rifle Teams." *Journal of Abnormal and Social Psychology*, LXV (1962), 365-375.
- McQuitty, L. L. "Isolating Predictor Patterns Associated with Major Criterion Patterns." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVII (1957), 3-42.
- Myers, T. I., Murphey, D., and Smith, S. "The Effect of Sensory Deprivation and Social Isolation on Self-Exposure to Propa-

- ganda and Attitude Change." *American Psychologist*, XVIII (1963), 440. (Abstract)
- Osler, S. F. and Trautman, G. E. "Concept Attainment: II Effect of Stimulus Complexity upon Concept Attainment at Two Levels of Intelligence." *Journal of Experimental Psychology*, LXII (1961), 9-13.
- Rasmussen, J. E. "An Experimental Approach to the Concept of Ego Identity as Related to Character Disorder." Unpublished Ph.D. thesis, American University, 1961.
- Saunders, D. R. "Moderator Variables in Prediction." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVI (1956), 209-222.
- Spielberger, C. D. "The Effects of Manifest Anxiety on the Academic Achievement of College Students." *Mental Hygiene*, in press.
- Taylor, J. A. and Spence, K. W. "The Relationship of Anxiety Level to Performance in Serial Learning." *Journal of Experimental Psychology*, XLIV (1952), 61-64.
- Torrance, E. P. and Ziller, R. C. "Risk and Life Experience: Development of a Scale for Measuring Risk-Taking Tendencies." USAF WADC Technology Note, No. 57-23, 1957.
- Whitman, R. M., Trosman, H., and Koenig, R. "Clinical Assessment of Passive-Aggressive Personality." *American Medical Association Archives of Neurology and Psychiatry*, LXXII (1954), 540-549.

THE MULTIPLE-CHOICE TEST AS AN INSTRUMENT IN PERPETUATING FALSE CONCEPTS

RALPH C. PRESTON
University of Pennsylvania

THIS study explored the assumption underlying the assertion by Skinner (1961): "Every wrong answer on a multiple-choice test increases the probability that a student will someday dredge out of his imperfect memory the wrong answer instead of the right one." The converse of the assertion was also explored.

To eliminate the vagueness of "one day," the study was designed to test the influence of wrong multiple-choice answers (and right multiple-choice answers) upon concepts during the same hour. Whether or not similar results would be obtained if the tests were spaced farther apart cannot, of course, be predicted on the basis of this investigation.

Procedure

Three tests of vocabulary based on ten words from United States history, designated as Test A, Test B, and Test C, were employed. Each of the tests was used to measure understanding of the following words which were presented and defined with care by the United States history textbook used by the Ss: charter, secession, federal, impressment, embargo, imperialism, reparations, moratorium, arbitration, diplomacy.

Test A required that a definition be constructed in writing for each of the words. Test B presented the ten words in multiple-choice form. The first item of Test B, for example, read:

Charter

- an alliance
- a grant of rights
- any legal document
- an informal agreement

Test C was identical to Test A, the *Ss* again constructing in writing their own definitions of the ten words. The three tests were administered successively, first Test A, then Test B, finally Test C, during one sitting and during a single hour. There were no time limits, and all *Ss* were permitted to finish each test.

The tests were administered to 79 eleventh grade students (54 boys and 25 girls) who were members of classes in United States history. Their IQs (derived from Otis Quick-Scoring Ability Test) ranged from 91 to 136, $M = 115.4$, $s.d. = 9.61$).

The definitions constructed by the *Ss* on Tests A and C were rated on a 3-point scale. The criteria of the WISC vocabulary test were adapted as follows. Any recognized meaning of the word as used in United States history (including an accurate synonym, a major use, a definitive feature, or several minor features which cumulatively indicated understanding) was accorded a score of 2; a vague or incomplete meaning, an attribute of the word unrelated to its use in United States history, or an unelaborated example of the use of the word, was accorded a score of 1; a response which was wrong, empty, trivial, or impoverished was accorded a score of 0. After rating independently a sampling of definitions, the investigator and an associate conferred on disagreements and refined the criteria for rating of the remaining items.

A count was made, for each word, of each *S* who selected a wrong definition on Test B (the multiple-choice test) which bore no similarity to the *S*'s constructed definition on Test A; and who, on Test C, constructed a wrong definition patterned on the one selected in Test B. This sequence of responses which, on the basis of Skinner's assumption, would occur with a probability greater than chance, will be referred to as Sequence 1.

A count was made, for each word, of each *S* who responded on Tests A and B according to Sequence 1, but whose response to Test C was *not* patterned on the wrong definition selected in Test B. One would expect, on the basis of Skinner's assumption, that this sequence (Sequence 2) would occur with a probability no greater than chance.

To test the converse of Skinner's proposition, a count was made, for each word, of each *S* whose constructed definition in Test A was wrong or partly wrong, whose selection of a definition in Test B was correct, and whose constructed definition in Test C was also correct,

TABLE 1
Examples of Four Types of Sequence in Defining Words on Tests A, B, and C

Word	Sequence 1 (conforms to Skinner's assumption)	Sequence 2 (deviates from Skinner's assumption)	Sequence 3 (conforms to converse of Skinner's assumption)	Sequence 4 (deviates from converse of Skinner's assumption)
Test A (constructed definition)	moratorium "A place of death such as a funeral parlor"	moratorium "A bad act such as treason"	imperialism "Government by a small group out for them- selves"	imperialism "A government ruled by one man"
Test B (multiple-choice selection)	"a threat"	"a regret"	"extending sovereignty"	"extending sovereignty"
Test C (constructed definition)	"a threat to another country"	"A bad act in congress"	"Extension of sov- eignty"	"a pure democrat"
Frequency of type of sequence	163	79	221	106
Percent of all sequences	21	10	28	13

and patterned after the definition selected in Test B. This sequence, the converse of Skinner's assumption, will be referred to as Sequence 3.

A count was made, for each word, of each *S* who, on Test A and B, responded as in Sequence 3, but whose constructed definition in Test C was wrong or partially wrong, bearing no relationship to the selection in Test B. This sequence will be referred to as Sequence 4.

Examples of the foregoing sequences appear in Table 1.

Results

Sequence 1 was followed more frequently than Sequence 2 for 8 of the 10 words, significantly more frequently for 3 of the words. (See Table 2.) Hence, Skinner's assumption receives partial support from these data.

Sequence 3 was followed more frequently than Sequence 4 for 9 of the 10 words, significantly more for 5 of the words. (See Table 2.) Thus, the converse of Skinner's assumption also receives support.

A study was made of sex and intelligence differences in the tendency to be influenced by wrong and by right multiple-choice answers. Each *S* was assigned a score representing his tendency to follow the sequence which supports Skinner's assumption computed as follows: $X = N_1 / (N_1 + N_2)$ in which N_1 is the frequency with which his answers followed Sequence 1 and N_2 is the frequency with which his answers followed Sequence 2. Each *S*'s tendency to follow the sequence which supports the converse of Skinner's assumption was computed through employment of the formula $X^1 = N_3 / (N_3 + N_4)$ in which N_3 and N_4 stand for the frequencies with which his answers followed Sequences 3 and 4, respectively. With respect to sex differences, mean X was the same for men as for women, and the difference between mean X^1 for men and mean X^1 for women was $-.01$ ($t = .16$, $P > .80$). Both X and X^1 correlated significantly with IQ: $r_x = .33$ ($P < .001$) and $r_{x^1} = .30$ ($P < .01$).

Conclusion

The conditioning effect of wrong selections of multiple-choice items was demonstrated for some words but not for others. Wrong selections conditioned the *Ss* less frequently than did right selections. The act of responding to a multiple-choice test item conditioned the ensuing response most of the time, whether the condition-

TABLE 2
Distribution of Four Types of Sequence in Defining Words on Tests A, B, and C

words	Multiple-choice selection (Test B) contained wrong concept not contained in prior constructed definition (Test A). Subsequent constructed definition (Test C):	was wrong and patterned on the wrong multiple-choice selection (Sequence 1)	was either right or wrong and <i>not</i> patterned on the wrong multiple-choice selection (Sequence 2)	d	X ²	Multiple-choice selection (Test B) was correct, whereas prior constructed definition (Test A) was not. Subsequent constructed definition (Test C):	was patterned on the correct multiple-choice selection (Sequence 3)	was <i>not</i> patterned on the correct multiple-choice selection (Sequence 4)	d	X ²
charter	15	7	1	8	2.91		18	9	9	3.00
secession	—	13	1	-1	1.00		19	2	17	13.76*
federal	43	13	13	30	16.07*		14	9	5	1.09
impressment	10	9	9	1	0.05		16	17	-1	0.03
embargo	2	4	4	-2	0.67		32	6	26	17.79*
imperialism	12	2	2	10	7.14*		29	23	6	0.69
reparations	13	9	9	4	0.73		23	7	16	8.53*
moratorium	41	13	13	28	14.52*		18	5	13	7.35*
arbitration	13	11	11	2	0.17		30	15	15	5.00*
diplomacy	14	10	10	4	0.67		22	13	9	2.31

* Significant at .05 level.

ing was factually misleading or not. Those of greater intelligence tended to be the more susceptible to such conditioning.

REFERENCE

Skinner, B. F. "Teaching Machines." *Scientific American*, CCV (1961), 90-102.

THE USE OF COMPLEX ALTERNATIVES IN MULTIPLE CHOICE ITEMS¹

HERBERT H. HUGHES

AND

W. EUGENE TRIMBLE

Colorado State College

In the past a number of variations of the multiple-choice item have been considered in an effort to obtain from a particular item more information about a student's knowledge. In conventionally scored multiple-choice items, complete information leads to a score of 1 and misinformation leads to a score of 0. Partial information may lead to a score of either 1 or 0. A realistic, although complex, picture of the problem is presented by Little (1962). He describes six combinations of information which might determine scores on an item:

1. positive correct information leading to a correct answer;
2. partial information leading to a correct answer;
3. total lack of information, selections by pure guess;
4. partial information leading to an incorrect answer;
5. positive misinformation leading to an incorrect answer;
6. a feeling that the test question is stupid, he knows the answer that is wanted, but does not agree and will not give the correct answer.

In any framework, however, assessing the knowledge of a student with partial, but less than complete, information on an item remains as the real enigma.

¹ Paper presented at joint session of National Council of Measurements in Education and American Educational Research Association, February, 1963, in Chicago.

Several investigators (Archer, 1962; Coombs, Milholland, and Womer, 1956; Dressel and Schmid, 1953) have evaluated the effect different methods of responding have on a student's opportunity to display varying degrees of information on multiple-choice items. Modified scoring weights (Davis, 1959; Merwin, 1959), various option sequences (Anderson, 1952; Mosier, Myers, and Price, 1945), and systematic reduction of the least effective distractors (Williams and Ebel, 1957), have been other ways sought to increase the effectiveness of multiple-choice items. Still another line of attack, the use of "none of these" and "Right answer not given" as alternatives, has been explored. In a study by Boynton (1950), inclusion of the "none of these" alternative made spelling items more difficult, and at the same time more effective in their discrimination. Neither Wesman and Bennett (1946) nor Rimland (1960) found overwhelming evidence to support the superiority of "none of these" or "Right answer not given" as alternatives, but "none of these" was very effective in certain items in Wesman and Bennett's study.

The hypothesis that the alternative "none of these" will thwart test sophisticated students who parlay partial information into correct answers is very appealing. The evidence on this point, however, is limited probably due in part to the fact that research is exceedingly difficult to conduct in this area.

Problem

The present study was concerned with the more general case of the "none of these" type of alternative, called in this study the complex alternative. Three types of complex alternatives were used, "All of the above are correct," "None of the above are correct," and "Both 1 and 2 above are correct" or a similar alternative combination. These alternatives may have several potential advantages. It is possible that they can increase, when this is desirable, the homogeneity and thus the difficulty of items. An alternative such as "Both 1 and 2 above are correct," when it is the correct answer, may be more effective than conventional alternatives if the goal of an item is to require the student to know two or more pieces of information before passing an item. It would also appear that complex alternatives could be of considerable value in increasing the level of discrimination and/or reducing the factor of guessing when three or fewer satisfactory conventional alternatives are available.

The purpose of the present study was to explore the potential of complex alternatives in tests of classroom achievement at the college level. On the basis of a pilot study it was hypothesized that items with complex alternatives would not only be more difficult, as reflected in both total scores and item difficulty, but would also permit greater discrimination by inducing a wider spread of scores than the conventional test items. Also predicted was a significantly higher relationship between the tests with modified items and two tests, one an open-response and the other developed by Coombs and his associates (1956), both designed to permit evaluation of degrees of information on an item.

Method

Both the experimental and the control tests for the study were constructed from carefully modified items from a test-item booklet and instructor written original items, both of which had been evaluated in previous classes. All tests consisted of 55 standard multiple-choice items. Approximately half the items were designed as purely factual items and half, insofar as was possible, as items to test concepts or principles. Every effort was made to sample equally from each of the sections covered in the course.

The test administered to the experimental and the control groups in the first examination consisted of 55 four-choice items. Similar tests were administered to the control group throughout the study. A standard, but admittedly arbitrary, plan which evolved out of the earlier pilot study was used to develop the experimental tests which were introduced in the second examination. Under the plan the control test for the second and third examinations was constructed first and then modified to form the two experimental tests.

Three items in every 10 of the first 50 items were not altered in the experimental tests. Thus, 15 unchanged items were available for control purposes in all of the tests. Six items in every 10 of the first 50 items, and the last five items, were modified by adding a form of the distractor "Both 1 and 2 above are correct" as a fifth choice on one experimental test form. The distractors "All of the above are correct" and "None of the above are correct" were alternated as fifth choices in the same items on the other experimental test form. The original correct answer was not changed in these modified items, thus providing 35 items which differed in only one respect between

the control and experimental test forms. The only difference in experimental and control test forms, therefore, was that in some items one extra choice was available, a crucial decision which offered the possibility of confounding the influence of complex alternatives and adding an extra choice.

In order to insure that the students taking the experimental tests would not automatically consider the additional choices as foils, one item in every 10 of the first 50 items was modified to make the fifth choice, a complex alternative, the correct answer. Since a total of 40 items in the experimental tests contained a complex alternative as a fifth choice, one in every eight (or a total of five) of these items was correct. Some of these five items required extensive changes which precluded any fruitful comparisons with the comparable items in the control test.

The choice of which items in each set of 10 should serve as controls and which should be modified in what ways was made on a subjective basis. This undoubtedly limited the scope of the study to special kinds of items, but any procedure artificially forcing a complex alternative on items would not resemble typical methods of test construction.

Each revised experimental test contained: 15 control items; 35 items (hereafter called comparison items) modified by adding as a fifth choice a complex alternative, "Both 1 and 2 above are correct" in one experimental test form and "All of the above are correct" or "None of the above are correct" in the other; 5 items modified to make the fifth choice, a complex alternative, the correct answer.

Subjects and Procedure

The subjects were 63 undergraduate students (24 men and 39 women) enrolled in a required freshman course in General Psychology. The class was divided randomly into three groups of 21 before the first examination.

Three examinations were given during the one quarter course. In the first examination all subjects received the same 55 item 4-choice test. In the second examination, group I (the control) received the standard 4-choice tests, group II the tests containing the complex alternatives "Both 1 and 2 above are correct," and group III the tests containing the alternatives "All of the above are correct" or "None of the above are correct." The subjects had been informed

that the tests would vary in format and that for grading purposes three separate distributions would be established on each test. It was pointed out that the instructor had used both kinds of items in previous courses and was interested in knowing more about their relative value.

The class met before each examination in one classroom and proceeded to a larger room where individual tests had been laid out on previously assigned desks. This permitted standard starting times for all groups.

The procedure for the first stage of the third examination was identical to the second testing session. Immediately following the regular examination period, all three groups responded to the control test (without complex alternatives) using Coomb's method in which the subjects mark the *distractors* and *not* the correct answers. After a 45 minute break for supper the three groups responded to the control test again using an open-response method in which the subjects wrote the reason every alternative in each item was or was not correct. The subjects were not informed that they would be responding to the same test items each time until the directions for each new method were read. Detailed directions, built around concrete examples, were used for the Coombs and the open-response tests. The purpose of the open-response test was to establish a criterion measure of the degree of information the subjects might have. The Coombs method was scored by his procedures. The following method was used to score the open-response test.

Point value of 2 = Knew the answer, by partial or complete information.

Point value of 1 = Didn't know answer, but had partial information.

Point value of 0 = Didn't know answer, and didn't have even partial information.

Results

Inspection of Tables 1, 2, and 3 indicates that there were no significant differences between the control and the experimental groups on the *control* items throughout the study. The same was true of the significance tests of the item difficulty for the three forms of tests. Reliability coefficients, estimated by Kuder-Richardson formula 21, remained relatively consistent throughout the study. Spearman-

TABLE 1

*Means, Standard Deviations, Reliability Coefficients, and Analysis of Variance
(First Testing)*

First Testing—all groups administered same test all 55 items included in analyses			
	<i>M</i>	<i>SD</i>	<i>K-R₂₁</i>
Group I (Control)	38.95	5.21	.59
Group II (Experimental)	38.29	6.96	.78
Group III (Experimental)	38.52	5.66	.65
<i>Analysis of Variance</i>			
<i>F</i> = .06			
<i>F</i> _{max} (homogeneity) = 1.79			

Brown and K-R 20 coefficients were also computed for all groups on the first testing; since there was high agreement among the three coefficients, and the K-R 21 normally underestimates the K-R 20, only the K-R 21 was used in subsequent analyses.

On the second and third testing significant differences were found on the 35 comparison items. Further evaluation of these differences confirmed the first hypothesis. Dunn's test, in which the control mean is compared against each of the two experimental means, showed mean scores were significantly lower in both experimental groups on the second examination, and in experimental group II (whose tests contained "Both 1 and 2" alternatives) on the third examination. These results were substantiated by analysis of the item difficulties. Ten of the 35 "Both 1 and 2" items and 9 of the 35 "None of the above" or "All of the above" items were significantly

TABLE 2

*Means, Standard Deviations, Reliability Coefficients, and Analyses of Variance
(Second Testing)*

Second Testing—analyses on 15 control and 35 comparison items			
	<i>M</i>	<i>SD</i>	<i>K-R₂₁</i>
Group I (Control)	27.19	4.04	.65
Group II (Experimental—Both)	22.05	5.15	.71
Group III (Experimental—None of the above)	23.76	4.29	.61
<i>Analyses of Variance</i>			
comparison items	<i>F</i> = 7.05**		
control items	<i>F</i> = .43		
<i>F</i> _{max} (homogeneity)	= 1.62		
Dunn—I with II	<i>c</i> = 3.60*		
I with III	<i>c</i> = 2.40*		

* Significant at .05 level.

** Significant at .01 level.

TABLE 3

*Means, Standard Deviations, Correlations, Reliability Coefficients,
and Analyses of Variance (Third Testing)*

Third Testing—analyses on 15 control and 35 comparison items			
	<i>M</i>	<i>SD</i>	<i>K-R²¹</i>
Group I (Control)	24.38	4.17	.59
Group II (Experimental—Both)	20.10	5.61	.71
Group III (Experimental—None of the above)	21.05	5.11	.70
<i>Analyses of Variance</i>			
comparison items	$F = 4.27^*$	Dunn—I with II $c = 2.71^*$	
control items	$F = .31$	I with III $c = 2.11$	
F_{\max} (homogeneity)	$= 1.81$		
<i>Correlations with Coombs and Open-Response</i>			
Coombs with I	$r = .89$	Open-Response with I $r = .83$	
with II	$r = .71$	with II $r = .78$	
with III	$r = .70$	with III $r = .66$	

* Significant at .05 level.

more difficult ($P = .05$) than the conventional items in the second testing. In the third testing in both experimental tests 8 of 35 comparison items were significantly more difficult than their conventional counterparts on the control test. In both cases this was a significantly greater proportion than was found in the 35 comparison items in the control tests. Additional inspection revealed that in the second testing 16 of the 19 items were made significantly more difficult by adding a complex alternative and had moved into the 35-65 percent range of difficulty. In the third testing 14 of the 16 items did likewise.

The hypothesis that each experimental test would have a significantly greater variation than the control test, which was in effect a prediction of heterogeneity of variance, was not supported. (See Tables 2 and 3.)

TABLE 4

Group Comparisons on Coombs and Open-Response Tests (Third Testing)

	<i>Coombs</i>			<i>Open-Response</i>		
	<i>M</i>	<i>SD</i>	<i>K-R²¹</i>	<i>M</i>	<i>SD</i>	<i>Sp-Br</i>
Group I (Control)	22.95	3.63	.78	41.29	10.72	.91
Group II (Experimental)	22.81	4.38	.85	38.38	9.30	.90
Group III (Experimental)	21.81	4.07	.78	39.19	8.53	.88
<i>Analyses of Variance</i>						
Comparison—Coombs				$F = .48$		
Control—Coombs				$F = .59$		
Comparison—Open-Response				$F = .49$		
Control—Open-Response				$F = .89$		

The correlation of each group's scores from test form I on the third testing with first their scores on the Coomb's test and then their scores on the open-response test failed to confirm the remaining hypothesis. In fact, the predicted order of correlation coefficients was reversed. The highest correlations occurred between the conventional test and the Coombs method, and the conventional test and the open-response test. (See Table 3.)

When approximately the upper 27 percent of scores was identified in each group on the second and third testings, with but two exceptions the same students appeared in the upper groups. A similar, but weaker trend, was true using the lower 27 percent of the scores from the same sets of tests. Mean differences between the upper 27 percent's and the lower 27 percent's revealed no clearcut evidence for superior discriminating power of either complex or conventional alternatives. Small numbers in each of the groups, however, prevented a precise evaluation of item discrimination with traditional procedures.

Approaching the analysis of the items from still another direction indicated that in both experimental groups on the final two tests a significantly higher proportion of the total number of comparison items marked incorrectly were missed by choosing the "Both 1 and 2," "None of the above," and "All of the above" alternatives. In addition, a significantly greater number of complex alternatives were chosen by both upper 27 percent and lower 27 percent groups than would have been expected.

Discussion and Conclusions

It appears that complex alternatives, particularly "Both 1 and 2 above are correct," can increase item difficulty. As distractors, they may have differential effects on the upper and lower 27 percent of a score distribution. What this does to item discrimination is still not clear. With other factors held constant moving an item into the 35-65 percent range of difficulty theoretically should increase the potential discriminating power of items and as a consequence their ability to evaluate degrees of information. No such evidence was detected in this study; furthermore, whether items containing complex alternatives *reward* or *penalize* students with greater degrees of information could not be determined from these data.

An important consideration in studies of this nature is how to

assess the validity of an item. When a significant difference between two items is obtained, which of the items is rewarding the student with greater knowledge? Content validity, based on an internal analysis of items, won't answer this. The hope that the Coomb's and open-response methods, used as external criterion measures, would help to assess validity was not realized in this study.

There are a number of possible reasons for failing to ferret out the effects of the complex alternatives. The small number of items and subjects used seems to be the most crucial. The many alternative procedures and designs available in this type of study, the role that guessing behavior may play (if students at the lower end of the score distribution are already guessing, changes which occur over the remaining part of the distribution may be obscured), and the differential effect that a complex alternative may have from item to item are among the possible reasons.

Several problems warrant further attention. With adjustments in the procedure and design based on the previous discussion, a study of the effect of adding the alternatives "Both 1 and 2" and "Neither 1 or 2" to items with two well written original choices should be fruitful; if successful, this would be a blessing to classroom teachers who are hard pressed to find enough distractors for multiple-choice items. External validity needs more concentrated study in cases where the ability of an item to measure degrees of knowledge is at stake. The Coombs and the open-response methods may hold promise, and a larger study patterned after the present one might reveal their effectiveness. One thing is certain. As long as test constructors, neophyte or experienced, continue to use great flexibility in item writing, it seems incumbent upon the researcher to determine the consequences. One closing comment—at the end of the class session in which the students were first oriented to the use of complex alternatives as answers, one very enlightened young lady came rushing up to the instructor and blurted out "So that's what they had 'none of these' on those college entrance test questions for—I *could've* picked it as one of the answers." In this age of complex theories and sophisticated approaches to test behavior, revelations like this make one stop and think.

REFERENCES

- Anderson, Scarvia. "Sequence in Multiple Choice Item Options." *Journal of Educational Psychology*, XLIII (1952), 364-368.

- Archer, S. N. "A Comparison of the Conventional and Two Modified Procedures for Responding to Multiple-Choice Test Items with Respect to Test Reliability, Validity, and Item Characteristics." Paper read at AERA Session on Measurement Studies, Atlantic City, February, 1962.
- Boynton, Marcia. "Inclusion of 'None of These' Makes Spelling Items More Difficult." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, X (1950), 431-432.
- Coombs, C. H., Milholland, J. E., and Womer, F. B. "The Assessment of Partial Knowledge." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVI (1956), 13-37.
- Davis, F. B. "Estimation and Use of Scoring Weights for Each Choice in Multiple-Choice Test Items." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIX (1959), 291-298.
- Dressel, P. L. and Schmid, J. "Some Modifications of the Multiple-Choice Item." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIII (1953), 574-595.
- Little, E. B. "Overcorrection for Guessing in Multiple-Choice Test Scoring." *Journal of Educational Research*, LV (1962), 245-252.
- Merwin, J. C. "Rational and Mathematical Relationships of Six Scoring Procedures Applicable to Three-Choice Items." *Journal of Educational Psychology*, L (1959), 153-161.
- Mosier, C. I., Myers, M. E., and Price, Helen G. "Suggestions for the Construction of Multiple-Choice Test Items." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, V (1945), 261-271.
- Rimland, B. "The Effects of Varying Time Limits and of 'Using Right Answer Not Given' in Experimental Forms of the U. S. Navy Arithmetic Test." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XX (1960), 533-539.
- Wesman, A. G. and Bennett, G. K. "The Use of 'None of These' as an Option in Test Construction." *Journal of Educational Psychology*, XXXVII (1946), 541-549.
- Williams, B. J. and Ebel, R. L. "The Effect of Varying the Number of Alternatives Per Item on Multiple-Choice Vocabulary Test Items." In *The Fourteenth Yearbook of the National Council on Measurements Used in Education*, The Council, 1957, 63-65.

THE PROBABILITY OF CHANCE SUCCESS ON OBJECTIVE TEST ITEMS

LEWIS R. AIKEN, JR.

The University of North Carolina at Greensboro

ALTHOUGH the usefulness of corrections for guessing on true-false and multiple-choice tests may be debatable (*e.g.*, see Lyster, 1951; Stanley, 1954; Jackson, 1955), such procedures are widely used. Several formulas for adjusting test scores for chance success have been suggested (see Gulliksen, 1950), but the following method is most frequently mentioned. A corrected total score for the examinee is calculated by subtracting from the number of correct answers¹ he gives $1/(n - 1)$ of the items which he marks incorrectly, n being the number of answer options. The use of the fraction $1/(n - 1)$ assumes, of course, that all n answer options on every item are equally plausible to the person who does not know the correct answer. Given this assumption, the method of correcting for guessing becomes quite general and need not be restricted to true-false or multiple-choice tests.

It is the purpose of this paper to describe a method for determining the probability of chance success on any type of objective test item. Such probabilities may then be used in correcting for guessing, in helping to determine the item and test difficulty and even in

¹ A suggested method for obtaining a score, uncorrected for guessing, for rearrangement items is to compute $P = (n_r^2 - j)/2$, where n_r is the number of sub-items to be rearranged and $j = 0$ if n_r is even or $j = 1$ if n_r is odd. Then the raw score, $S = P - \sum_{i=1}^{n_r} |d_i|$, is computed, where d_i is the difference between the correct numerical order for a sub-item and the numerical order assigned to it by the examinee. If it is desirable to weight the rearrangement item by the number of sub-items which comprise it in order to have a perfect score of n_r , S must be multiplied by $2/n_r$ if n_r is even or by $2n_r/(n_r^2 - 1)$ if n_r is odd. The examiner may then wish to subtract a correction for guessing from the weighted S value to obtain a final score for the examinee on the item.

setting grade boundaries. In addition, this general procedure for computing probabilities may be found to be applicable in areas other than testing.

Items as Stimulus-Response Matchings

In actuality, all objective test item types—true-false, multiple choice, completion, matching, rearrangement, etc.—may be considered as varieties of the matching item. They all give the examinee the task of matching one set of numbers, words, or groups of words to another set of numbers, words, or groups of words. For convenience, one of these sets will be referred to as responses (r) and the other as stimuli (s), although which set is r and which s is rather arbitrary. However, in the development of the general formula below, it will be assumed that $n_r \geq n_s$, where n_r is the number of response options and n_s the number of stimulus options. Thus, for the true-false item type, $n_r = 2$, $n_s = 1$. For the multiple-choice item, in the usual case $n_r = 3, 4$, or 5 and $n_s = 1$. For the completion item, $n_s = 1$ but n_r is undetermined, being something less than infinity. For the matching item, $n_s = 10$ to 20 and $n_r = 10+$ to $20+$ in the usual case. And for the rearrangement item, $n_s = n_r = 10$ to 20 the responses in this case being the rank orders or positions into which the n_s stimuli are arranged.

Development of the Formula

Using this response-stimulus paradigm of test items, the development of a formula for determining the probability of obtaining any possible number of correct $r \rightarrow s$ matches proceeds as follows. Let

n_s = the number of stimulus options in the item,

n_r = the number of response options in the item,

$s_a, s_b, s_c, s_d \dots$ = arbitrary stimuli, i.e. the a -th, b -th, c -th, d -th, \dots ,

$r_a, r_b, r_c, r_d \dots$ = the correct response matches for the respective arbitrary stimuli above,

$C(n_s, n_r)$ = the combinations of n_s things taken n_r at the time,

$Per(n_s, n_r)$ = the permutations of n_s things taken n_r at the time.

It is assumed that there is only one correct stimulus match for n_s responses, the remaining $n_r - n_s$ responses having no stimulus matches. It can be seen from the above definitions that the probability of r_a , the correct response to s_a , being matched to s_a is:

$$\Pr(r_a \rightarrow s_a) = \Pr(r_b \rightarrow s_b) = \Pr(r_c \rightarrow s_c)$$

$$= \dots = \Pr(r_s \rightarrow s_s) = \frac{\text{Per}(n_r - 1, n_s - 1)}{\text{Per}(n_r, n_s)} \quad (1)$$

The probability of $r_a \rightarrow s_a \wedge r_b \rightarrow s_b$ is:

$$\begin{aligned} \Pr(r_a \rightarrow s_a \wedge r_b \rightarrow s_b) &= \Pr(r_a \rightarrow s_a \mid r_b \rightarrow s_b) \cdot \Pr(r_b \rightarrow s_b) \\ &= \frac{\text{Per}(n_r - 2, n_s - 2)}{\text{Per}(n_r, n_s)}. \end{aligned}$$

In general, it can be shown that the joint probability of $r_a \rightarrow s_a, r_b \rightarrow s_b, r_c \rightarrow s_c, r_d \rightarrow s_d, \dots, r_z \rightarrow s_z$, where z represents any number of matches, is:

$$\begin{aligned} \Pr(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_z \rightarrow s_z) \\ &= \Pr(r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_z \rightarrow s_z) \\ &\quad \cdot \Pr(r_a \rightarrow s_a \mid r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_z \rightarrow s_z) \\ &= \frac{\text{Per}(n_r - n_m, n_s - n_m)}{\text{Per}(n_r, n_s)}, \quad (2) \end{aligned}$$

where n_m is the number of matches, the joint probability of which is desired.

The probability of $r_a \rightarrow s_a$ or $r_b \rightarrow s_b$ is:

$$\begin{aligned} \Pr(r_a \rightarrow s_a \vee r_b \rightarrow s_b) \\ &= \Pr(r_a \rightarrow s_a) + \Pr(r_b \rightarrow s_b) - \Pr(r_a \rightarrow s_a \wedge r_b \rightarrow s_b) \\ &= \frac{\text{Per}(n_r - 1, n_s - 1)}{\text{Per}(n_r, n_s)} + \frac{\text{Per}(n_r - 1, n_s - 1)}{\text{Per}(n_r, n_s)} \\ &\quad - \frac{\text{Per}(n_r - 2, n_s - 2)}{\text{Per}(n_r, n_s)} \\ &= \frac{2 \cdot \text{Per}(n_r - 1, n_s - 1)}{\text{Per}(n_r, n_s)} - \frac{\text{Per}(n_r - 2, n_s - 2)}{\text{Per}(n_r, n_s)}. \end{aligned}$$

In general, it can be shown that the probability of $r_a \rightarrow s_a$ or $r_b \rightarrow s_b$ or $r_c \rightarrow s_c$ or $r_d \rightarrow s_d$ or \dots or $r_z \rightarrow s_z$ is:

$$\begin{aligned} \Pr(r_a \rightarrow s_a \vee r_b \rightarrow s_b \vee r_c \rightarrow s_c \vee r_d \rightarrow s_d \vee \dots \vee r_z \rightarrow s_z) \\ &= \frac{n_s \cdot \text{Per}(n_r - 1, n_s - 1)}{\text{Per}(n_r, n_s)} - \frac{n_s(n_s - 1) \text{Per}(n_r - 2, n_s - 2)}{2! \text{Per}(n_r, n_s)} \\ &\quad + \frac{n_s(n_s - 1)(n_s - 2) \text{Per}(n_r - 3, n_s - 3)}{3! \text{Per}(n_r, n_s)} - \dots \\ &\pm \frac{n_s(n_s - 1)(n_s - 2) \dots (1) \text{Per}(n_r - n_s, 0)}{n_s! \text{Per}(n_r, n_s)} = \Pr(n_R \geq 1), \end{aligned}$$

where n_R should be read as "the number right."

Simplifying, we have

$$\begin{aligned} \Pr (n_R \geq 1) &= \frac{C(n_s, 1) \text{Per}(n_r - 1, n_s - 1)}{\text{Per}(n_r, n_s)} \\ &\quad - \frac{C(n_s, 2) \text{Per}(n_r - 2, n_s - 2)}{\text{Per}(n_r, n_s)} + \frac{C(n_s, 3) \text{Per}(n_r - 3, n_s - 3)}{\text{Per}(n_r, n_s)} \\ &\quad - \dots \pm \dots \pm \frac{C(n_s, n_s) \text{Per}(n_r - n_s, 0)}{\text{Per}(n_r, n_s)}, \end{aligned}$$

or, simplifying still further

$$\begin{aligned} \Pr (n_R \geq 1) &= \frac{C(n_s, 1)}{\text{Per}(n_r, 1)} - \frac{C(n_s, 2)}{\text{Per}(n_r, 2)} \\ &\quad + \frac{C(n_s, 3)}{\text{Per}(n_r, 3)} - \dots \pm \frac{C(n_s, n_s)}{\text{Per}(n_r, n_s)}, \end{aligned}$$

or

$$\Pr (n_R \geq 1) = \sum_{n_s=1}^{n_s} (-1)^{n_s-1} \frac{C(n_s, n_s)}{\text{Per}(n_r, n_s)}.$$

Also, it can be shown that

$$\begin{aligned} \Pr [(r_s \rightarrow s_s \wedge r_t \rightarrow s_t) \vee (r_s \rightarrow s_s \wedge r_e \rightarrow s_e) \\ \vee (r_t \rightarrow s_t \wedge r_e \rightarrow s_e) \vee \dots \vee (r_s \rightarrow s_s \wedge r_e \rightarrow s_e)] \\ = \frac{n_s(n_s - 1) \text{Per}(n_r - 2, n_s - 2)}{2! \text{Per}(n_r, n_s)} \\ - \frac{2n_s(n_s - 1)(n_s - 2) \text{Per}(n_r - 3, n_s - 3)}{3! \text{Per}(n_r, n_s)} \\ + \frac{3n_s(n_s - 1)(n_s - 2)(n_s - 3) \text{Per}(n_r - 4, n_s - 4)}{4! \text{Per}(n_r, n_s)} \\ - \dots \pm \frac{(n_s - 1)n_s(n_s - 1) \dots (1) \text{Per}(n_r - n_s, 0)}{n_s! \text{Per}(n_r, n_s)}. \end{aligned}$$

Simplifying,

$$\begin{aligned} \Pr (n_R \geq 2) &= \frac{C(n_s, 2) \text{Per}(n_r - 2, n_s - 2)}{\text{Per}(n_r, n_s)} \\ &\quad - \frac{2C(n_s, 3) \text{Per}(n_r - 3, n_s - 3)}{\text{Per}(n_r, n_s)} + \frac{3C(n_s, 4) \text{Per}(n_r - 4, n_s - 4)}{\text{Per}(n_r, n_s)} \\ &\quad - \dots \pm \frac{(n_s - 1)C(n_s, n_s) \text{Per}(n_r - n_s, 0)}{\text{Per}(n_r, n_s)}. \end{aligned}$$

simplifying still further

$$\begin{aligned} \Pr (n_R \geq 2) &= \frac{C(n_s, 2)}{\text{Per}(n_r, 2)} - \frac{2C(n_s, 3)}{\text{Per}(n_r, 3)} \\ &\quad + \frac{3C(n_s, 4)}{\text{Per}(n_r, 4)} - \dots \pm \frac{(n_s - 1)C(n_s, n_s)}{\text{Per}(n_r, n_s)}, \\ \Pr (n_R \geq 2) &= \sum_{n_s=2}^{n_s} (-1)^{n_s-2} \frac{C(n_s - 1, n_s - 2)C(n_s, n_s)}{\text{Per}(n_r, n_s)}. \end{aligned} \quad (4)$$

In general, $\Pr (n_R \geq n_i)$, where n_i is any number greater than zero is found by

$$\begin{aligned} &[(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge \dots \wedge r_u \rightarrow s_u) \\ &\quad \vee (r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_v \rightarrow s_v) \\ &\quad \vee (r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge r_e \rightarrow s_e \wedge \dots \wedge r_w \rightarrow s_w) \vee \dots \text{etc.}], \\ &C(n_s, n_i) \text{ terms of } n_i \text{ matches each} \\ &C(n_s, n_i) \Pr (r_a \rightarrow s_a \wedge r_b \rightarrow s_b \\ &\quad \wedge r_c \rightarrow s_c \wedge \dots \wedge r_u \rightarrow s_u) \left. \vphantom{\begin{aligned} &C(n_s, n_i) \Pr (r_a \rightarrow s_a \wedge r_b \rightarrow s_b \\ &\quad \wedge r_c \rightarrow s_c \wedge \dots \wedge r_u \rightarrow s_u) \end{aligned}} \right\} 1(n_i) \text{ terms} \end{aligned}$$

$$\begin{aligned} &n_i C(n_s, n_i + 1) \\ &\Pr [(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge \dots \wedge r_v \rightarrow s_v) \\ &\quad (r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_w \rightarrow s_w)] \left. \vphantom{\begin{aligned} &\Pr [(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge \dots \wedge r_v \rightarrow s_v) \\ &\quad (r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_w \rightarrow s_w)] \end{aligned}} \right\} 2(n_i + 1) \text{ terms} \\ &\frac{n_i(n_i + 1)C(n_s, n_i + 2)}{2!} \end{aligned}$$

$$\begin{aligned} &\Pr [(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge \dots \wedge r_x \rightarrow s_x) \\ &\quad (r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_y \rightarrow s_y) \\ &\quad (r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge r_e \rightarrow s_e \wedge \dots \wedge r_z \rightarrow s_z)] \left. \vphantom{\begin{aligned} &\Pr [(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge \dots \wedge r_x \rightarrow s_x) \\ &\quad (r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge \dots \wedge r_y \rightarrow s_y) \\ &\quad (r_c \rightarrow s_c \wedge r_d \rightarrow s_d \wedge r_e \rightarrow s_e \wedge \dots \wedge r_z \rightarrow s_z)] \end{aligned}} \right\} 3(n_i + 2) \text{ terms} \end{aligned}$$

...

$$\begin{aligned} &\frac{n_i(n_i + 1)(n_i + 2) \dots (n_s - 1)}{(n_s - n_i)!} C(n_s, n_s) \\ &\cdot \Pr [(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge \dots \wedge r_x \rightarrow s_x) \\ &\quad \wedge (r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge \dots \wedge r_y \rightarrow s_y) \\ &\quad \wedge \dots \wedge \dots \wedge \dots \wedge \dots \wedge \dots \\ &\quad \wedge (r_y \rightarrow s_y \wedge r_z \rightarrow s_z \wedge \dots \wedge r_x \rightarrow s_x)] \left. \vphantom{\begin{aligned} &\Pr [(r_a \rightarrow s_a \wedge r_b \rightarrow s_b \wedge \dots \wedge r_x \rightarrow s_x) \\ &\quad \wedge (r_b \rightarrow s_b \wedge r_c \rightarrow s_c \wedge \dots \wedge r_y \rightarrow s_y) \\ &\quad \wedge \dots \wedge \dots \wedge \dots \wedge \dots \wedge \dots \\ &\quad \wedge (r_y \rightarrow s_y \wedge r_z \rightarrow s_z \wedge \dots \wedge r_x \rightarrow s_x)] \end{aligned}} \right\} n_s(n_s - n_i + 1) \text{ terms.} \end{aligned}$$

So,

$$\begin{aligned}
 \Pr(n_R \geq n_i) &= \frac{C(n_s, n_i) \text{Per}(n_r - n_i, n_s - n_i)}{\text{Per}(n_r, n_s)} \\
 &- \frac{n_i C(n_s, n_i + 1) \text{Per}(n_r - n_i - 1, n_s - n_i - 1)}{\text{Per}(n_r, n_s)} \\
 &+ \frac{n_i(n_i + 1) C(n_s, n_i + 2) \text{Per}(n_r - n_i - 2, n_s - n_i - 2)}{2! \text{Per}(n_r, n_s)} - \dots \pm \dots \\
 &\pm \frac{n_i(n_i + 1)(n_i + 2) \dots (n_s - 1) C(n_s, n_s) \text{Per}(n_i - n_s, 0)}{(n_s - n_i)! \text{Per}(n_r, n_s)} \\
 &= \frac{C(n_s, n_i)}{\text{Per}(n_r, n_i)} - \frac{n_i C(n_s, n_i + 1)}{1! \text{Per}(n_r, n_i + 1)} \\
 &+ \frac{n_i(n_i + 1) C(n_s, n_i + 2)}{2! \text{Per}(n_r, n_i + 2)} - \dots \\
 &\pm \dots \pm \frac{n_i(n_i + 1)(n_i + 2) \dots (n_s - 1) C(n_s, n_s)}{(n_s - n_i)! \text{Per}(n_r, n_s)}.
 \end{aligned}$$

If we use the symbol n_x as a variable ranging from n_i to n_s , then

$$\begin{aligned}
 \frac{n_i(n_i + 1)(n_i + 2) \dots (n_s - 1)}{(n_s - n_i)!} &= \frac{n_i(n_i + 1)(n_i + 2) \dots (n_x - 1)}{(n_x - n_i)!} \\
 &= \frac{(n_x - 1)!}{(n_i - 1)!(n_x - n_i)!} = C(n_x - 1, n_x - n_i).
 \end{aligned}$$

So the general formula becomes

$$\Pr(n_R \geq n_i) = \sum_{n_x=n_i}^{n_s} (-1)^{n_s-n_i} \frac{C(n_x - 1, n_x - n_i) C(n_s, n_x)}{\text{Per}(n_r, n_x)}. \quad (5)$$

Note that when $n_x = n_i$, $C(n_x - 1, n_x - n_i) = C(n_i - 1, 0) = 1$.

Using the Formula

Formula (5) may be used to determine the probability of the number correct by chance being inclusively greater than any value of $n_i \leq n_s$, with n_s and $n_r, n_r \geq n_s$, also equal to any values. Table 1 gives such probabilities of $n_R \geq 1, 2, 3, 4$, or 5 for values of n_s from 1 to 10 and n_r from 2 to 10.² The exclusive probability of any n_i

² As Table 1 shows, $\Pr(n_R \geq n_i)$ is essentially the same for a given value of n_i in all cases where $n_r = n_s > 4$. The reason is that formula (5) represents a convergent, alternating series, and the convergence is most rapid when $n_r = n_s$. Thus, in the special case where $n_r = n_s$, formula (5) may be reduced to:

$$(a) \quad \Pr(n_R \geq n_i) = \frac{1}{(n_i - 1)!} \sum_{j=0}^{(n_s - n_i)} (-1)^j \frac{1}{j!(n_i + j)},$$

TABLE 1

Chances out of 100 that the Number of Answers Right by Guessing (n_R) Will Be Greater than or Equal to Selected Magnitudes (n_i) for Various Numbers of Stimulus and Response Options

Number of Response Options	Number of Stimulus Options	n_i					Number of Response Options	Number of Stimulus Options	n_i				
		1	2	3	4	5			1	2	3	4	5
		Chances out of 100 that $n_R \geq n_i$							Chances out of 100 that $n_R \geq n_i$				
2	1	50					8	1	13				
2	2	50	50				8	2	23	02			
							8	3	32	05	00		
3	1	33					8	4	40	09	01	00	
3	2	50	17				8	5	47	13	02	00	00
3	3	67	17	17			8	6	53	17	04	01	00
							8	7	59	22	06	01	00
							8	8	63	26	08	02	00
4	1	25											
4	2	42	08				9	1	11				
4	3	54	17	04			9	2	21	01			
4	4	63	29	04	04		9	3	29	04	00		
							9	4	37	07	01	00	
5	1	20					9	5	43	10	02	00	00
5	2	35	05				9	6	49	14	03	00	00
5	3	47	11	02			9	7	55	18	04	01	00
5	4	56	19	04	01		9	8	59	22	06	01	01
5	5	63	26	09	01	01	9	9	63	26	08	02	01
6	1	17					10	1	10				
6	2	30	03				10	2	19	01			
6	3	41	08	01			10	3	27	03	00		
6	4	50	14	03	00		10	4	34	06	00	00	
6	5	57	20	05	01	00	10	5	40	09	00	00	00
6	6	63	27	08	02	00	10	6	46	12	03	00	00
							10	7	51	15	03	00	00
7	1	14					10	8	55	19	05	01	00
7	2	26	02				10	9	60	23	06	01	01
7	3	36	06	00			10	10	63	26	08	02	01
7	4	45	11	02	00								
7	5	52	16	03	00	00							
7	6	58	21	05	01	00							
7	7	63	26	08	02	00							

where the variable " j " = $n_R - n_i$ is introduced to simplify the expression. Then if we let $n_i = 1$, formula (a) becomes:

$$\begin{aligned}
 (b) \quad \Pr(n_R \geq 1) &= \sum_{j=0}^{n_R-1} (-1)^j \frac{1}{(j+1)!} \\
 &= 1 - 1/2! + 1/3! - \dots (-1)^{(n_R-1)} \frac{1}{n_R!} = 1 - e^{-1},
 \end{aligned}$$

which illustrates how formula (5) is related to the alternating, convergent e^{-1} series.

may be obtained simply by subtracting the inclusive $\Pr (n_R \geq n_i + 1)$ from the inclusive $\Pr (n_R \geq n_i)$, viz.:

$$\Pr (n_R = n_i) = \Pr (n_R \geq n_i) - \Pr (n_R \geq n_i + 1).$$

In this way, a table of exclusive probabilities may be formed from Table 1.

Summary

All objective test items are considered as varieties of the matching item, and a response-stimulus paradigm of items is described. A formula, from which the probability of obtaining any number of correct matches on a given item by chance may be computed, is presented, and a table of probabilities determined from the formula is given.

REFERENCES

- Gulliksen, H. *Theory of Mental Tests*. New York: John Wiley & Sons, 1950.
- Jackson, R. A. "Guessing and Test Performance." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XV (1955), 74-79.
- Lyerly, S. B. "A Note on Correcting for Chance Success in Objective Tests." *Psychometrika*. XVI (1951), 21-30.
- Stanley, J. C. "'Psychological' Correction for Chance." *Journal of Experimental Education*, XXII (1954), 297-298.

THE INFLUENCE ON TEST ANXIETY SCORES OF STRESSFUL VERSUS NEUTRAL CONDITIONS OF TEST ADMINISTRATION¹

CHARLES P. SMITH

Princeton University

In proposing criteria with which to evaluate measures of motivation, McClelland (1958) maintains that the measure of a motive should reflect its presence or absence or its variations in strength. He believes that this kind of sensitivity has been demonstrated for the thematic apperceptive measure of Need for Achievement (McClelland, Atkinson, Clark, and Lowell, 1953), since scores are higher under "achievement-oriented" conditions than under "relaxed" conditions. McClelland has doubts about self-report measures of motivation, however, and asks, "Do individuals score higher on any of the various questionnaire measures of anxiety as experimentally-induced anxiety increases" (1958, p. 24)?

In the present study, the sensitivity of the Test Anxiety Questionnaire (Mandler and Sarason, 1952) to situationally induced changes in motivation level is studied by comparing scores obtained under neutral and under stressful testing conditions. The comparability of "neutral" and "aroused" Test Anxiety scores is further investigated by comparing the relationships of both types of anxiety scores to scores on three criterion variables—intelligence, performance level (as measured by examination grades), and persistence (as measured by time spent in examinations).

¹ A preliminary version of this paper was presented at the Seventieth Annual Convention of the American Psychological Association, St. Louis, Missouri, August 31, 1962. This investigation was supported by a Public Health Service research grant (No. M-5301) from the National Institutes of Health, Public Health Service.

Method

Subjects

Subjects were 146 students in Introductory Psychology who were tested during regular class sessions.

Measure of Test Anxiety

Test Anxiety scores were obtained from the first section of the Test Anxiety Questionnaire which concerns reactions to group intelligence tests. Previous studies show correlations between scores from this section and the total questionnaire to be .90 (Litwin, 1958, $N = 50$), .84 (Mixon, in progress, $N = 175$), and .89 (Smith, 1961, $N = 215$). Each of the 12 anxiety items was scored as a five-point scale.

Measures of Criterion Variables

Intelligence was measured by 30 minutes of performance on the Otis Gamma group intelligence test (form Am). The test was introduced as a "nationally standardized general intelligence test." Subjects were asked to write their names on the test, and the timing of the test was fully apparent. *Level of performance* was measured by grades on multiple choice mid-term and final examinations. *Persistence* was measured by recording time taken by each subject to complete each exam.

Experimental Conditions

Subjects were divided into two groups for administration of the Test Anxiety Questionnaire and the Otis intelligence test. Since random assignment of subjects was not possible, each group was composed of alternate sections. In Group A ($N = 71$) subjects took the Test Anxiety Questionnaire under *neutral* conditions prior to taking the Otis test. That is, no instructions were given either to arouse or to relax motivation. Subjects in this group were unaware that an intelligence test would be given. In Group B ($N = 75$) subjects took the Test Anxiety Questionnaire immediately following the Otis test. No special instructions were given for the anxiety questionnaire, but it is assumed that anxiety was aroused during the intelligence test and carried over to the anxiety questionnaire.

*Results**Influence of Testing Conditions on Test Anxiety Scores*

Results concerning the influence on Test Anxiety scores of neutral versus stressful testing conditions are as follows: the mean neutral Test Anxiety score of 29.7 is slightly higher and not significantly different ($t = .57$, $df = 144$, n.s.) from the mean aroused Test Anxiety score of 28.8. Standard deviations of 8.5 and 8.9 in both groups are likewise nearly identical. In short no "arousal effect" is apparent. If anything Test Anxiety scores obtained following a presumably stressful intelligence test are slightly lower than neutral Test Anxiety scores.

Influence of Test Conditions on Relationship of Test Anxiety to Criterion Variables

Correlations between Test Anxiety and other variables are presented in Table 1. There is a negative relationship between Test Anxiety and intelligence in both Group A ($r = -.36$, $p < .05$) and Group B ($r = -.22$, $p < .10$). In both Groups A and B relationships between Test Anxiety and mid-term and final exam grades are small, nonsignificant, and negative ranging from $-.05$ to $-.11$. Relationships between Test Anxiety and time spent in examinations are positive in both Groups A and B and range from .24 ($p < .05$) to .38 ($p < .01$).

To determine whether different testing conditions influenced the relationship between Test Anxiety and the criterion variables, the magnitude of the correlations obtained in Groups A and B was compared. No significant differences were obtained. In other words, neutral and aroused Test Anxiety scores are related in the same way

TABLE 1

Correlations between Neutral and Aroused Test Anxiety Scores and Intelligence, Exam Scores, and Time Spent in Exams

Test Anxiety	N	Intelligence (Otis)	Mid-Term Exam	Final Exam	Time on Mid-Term	Time on Final
Neutral (Group A)	71	-.36**	-.05	-.06	.38**	.24**
Aroused (Group B)	75	-.22*	-.10	-.11	.26**	.30**

* $P < .10$ (two-tailed test).

** $P < .05$ (two-tailed test).

to intelligence, performance level, and persistence in the present study.²

Discussion

The results indicate that between neutral and stressful testing conditions, there are no differences in means or variances of Test Anxiety scores, and that both neutral and aroused Test Anxiety scores are related in the same way to intelligence, grades, and time spent in examinations.

One might ask why, in fact, Test Anxiety scores were not higher under stressful than under neutral conditions of testing. One possibility is that stressful testing conditions do indeed arouse higher anxiety, but that the heightened anxiety in turn arouses defensive processes which prevent the person from admitting his anxiety. This explanation is suggested by Sarason, et al. (1960) in their discussion of a "position effect" which they discovered. These authors found that scores on a Test Anxiety Scale for Children were lower when the scale followed rather than preceded a general anxiety scale. They suggest that the general anxiety scale may have produced defensiveness and made the children less willing to admit anxiety on a subsequent scale. The fact that Test Anxiety scores obtained following an intelligence test in the present study are slightly lower than Test Anxiety scores obtained under neutral conditions is consistent with this interpretation, but the difference is too small to represent clear support for this view.

It should also be noted that the nature of the sample may have

² Neutral and aroused Test Anxiety scores are related differently to Need for Achievement scores which were obtained under neutral conditions from some of the subjects in the present study earlier in the term for other purposes. For Group A, the correlation is .23 ($N = 59$), and for Group B, $-.15$ ($N = 64$). The difference is significant ($Z = 2.14$, $p < .05$). There are several reasons to think that this unexpected difference reflects chance variation in samples rather than the differential influence of testing conditions. First, similar differences, though not as great, have occurred previously. For example, correlations between neutral Test Anxiety scores and neutral Need for Achievement scores of .11 ($N = 125$, Smith, 1961) and $-.15$ ($N = 47$, Atkinson and Litwin, 1960) have been reported. Second, since the latter correlation is identical with that obtained in Group B, there is no strong reason to believe that the Test Anxiety scores in Group B were influenced by the stressful testing conditions for that group. Third, it is difficult to evaluate the effect of missing subjects on the correlations. Since Groups A and B were taught by different teachers, the 12 students who absented themselves from class in Group A, may be different from the 11 who failed to attend in Group B.

influenced the effect of the "arousal" conditions on the Test Anxiety scores, since it is conceivable that a group of highly intelligent students would not be greatly threatened by an intelligence test. Hence, the present procedure should be carried out for other populations before the generality of the results can be established. Also, students who felt, on completing the intelligence test, that they did well may have been relieved of their anxiety before taking the anxiety questionnaire. In a further study subjects might be asked how well they thought they did on the intelligence test prior to the administration of the anxiety questionnaire.

A more likely interpretation of the results, however, derives from the fact that the Test Anxiety Questionnaire *does not purport* to measure momentary changes in anxiety level in an individual, but rather is designed to measure his characteristic level of anxiety in certain types of testing situations. The latter possibility assumes that a person can accurately report the feelings he has had in test situations, whether or not his report has recently been preceded by the experience of taking a test. There is considerable evidence to support this assumption. In the present study, for example, one group of subjects took the first section of the Test Anxiety Questionnaire, which deals with reactions toward group intelligence tests, immediately after taking a group intelligence test. Having just taken an intelligence test should have produced maximally accurate self-reports, assuming no defensiveness, yet the scores of this group appear to be highly similar to Test Anxiety scores obtained under neutral conditions. (A further study comparing Test Anxiety scores of the *same* subjects before and after an unanticipated intelligence test might provide further information on this point, but would encounter the difficulties associated with test-retest designs.)

There are also other indications that subjects can report accurately on their emotional conditions during tests without having just taken a test. For example, Mandler and Sarason (1952) showed that observed manifestations of anxiety during a test, such as perspiration, excessive movement, and questioning of instructions, corresponded with classification of subjects as high and low in anxiety by means of Test Anxiety scores ($\Phi = .59, p = .001$). Similarly, Kissel and Littig (1962) report that subjects with high Test Anxiety have higher galvanic skin response scores than subjects with low Test Anxiety when working insoluble tasks. Their results suggest,

once more, that subjects can accurately report or recall their emotional states during testing conditions—degree of perspiration being one of the experiences asked about in the Test Anxiety Questionnaire.³

The results of the present study indicate that the Test Anxiety Questionnaire does not meet McClelland's (1958) criterion that a measure of motivation should reflect situational changes in an individual's level of motivation. This would appear to be a limitation of the measure insofar as it purports to measure only *one level* of a person's anxiety in *one sort of situation*. On the other hand, the insensitivity of the Test Anxiety Questionnaire to situational changes in testing conditions can be regarded as a definite asset, since there is some evidence (summarized by Smith, 1961) to indicate that the thematic apperceptive measure of Need for Achievement is overly sensitive to situational factors so that it is difficult to obtain comparable scores on two different groups of subjects.

Summary

Test Anxiety measured under neutral ($N = 71$) and aroused ($N = 75$) conditions is related to intelligence, exam grades, and time spent in exams. There are no differences between means and variances of Test Anxiety measured prior to (neutral) and following (aroused) a group intelligence test. Both neutral and aroused Test Anxiety scores are related in the same way to intelligence, grades, and persistence. The results suggest that the Test Anxiety Questionnaire measures characteristic level of anxiety in test situations under different conditions of administration, rather than momentary strength of anxiety at time of administration.

REFERENCES

- Atkinson, J. W. and Litwin, G. H. "Achievement Motive and Test Anxiety Conceived as Motive to Approach Success and Motive to Avoid Failure." *Journal of Abnormal and Social Psychology*, LX (1960), 52-63.
- Kissel, S. and Littig, L. W. "Test Anxiety and Skin Conductance." *Journal of Abnormal and Social Psychology*, LXV (1962), 276-278.

³It may be that students can report their exam experiences accurately because they have had many experiences with tests and because they tend to talk about and analyze these experiences. This apparent accuracy does not necessarily imply that reports of *infrequently* experienced situations of which persons tend to be less self conscious would be as accurate.

- Litwin, G. H. "Motives and Expectancy as Determinants of Preference for Degrees of Risk." Unpublished honors thesis, University of Michigan, 1958.
- McClelland, D. C. "Methods of Measuring Human Motivation." In J. W. Atkinson (Ed.), *Motives in Fantasy, Action, and Society*. Princeton: Van Nostrand, 1958, Pp. 7-42.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., and Lowell, E. L. *The Achievement Motive*. New York: Appleton-Century Crofts, 1953.
- Mandler, G. and Sarason, S. B. "A Study of Anxiety and Learning." *Journal of Abnormal and Social Psychology*, XLVII (1952), 166-173.
- Mixson, R. J. "Determinants of Academic Transfer and Withdrawal." Ph.D. dissertation in progress, Princeton University.
- Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., and Ruebush, B. K. *Anxiety in Elementary School Children*. New York: Wiley, 1960.
- Smith, C. P. "Situational Determinants of the Expression of Achievement Motivation in Thematic Apperception." Unpublished doctoral dissertation, University of Michigan, 1961.

THE CONCEPT OF ITEM DIFFICULTY IN PERSONALITY MEASUREMENT: A NOTE ON HANLEY'S FORMULATION

PAUL DEMPSEY
University of California, Davis

In the Autumn, 1962, issue of this journal, Hanley reported the results of a study of response latencies to personality items. The study design provides a neat demonstration of how to gather relevant experimental data with a minimum of instrumental controls, and Hanley was able to report a significant difference between the mean response latencies to MMPI items showing a moderate endorsement frequency (40-60 percent) and items showing either a low or a high frequency (0-20 percent or 80-100 percent). With these results there can be no quarrel. In providing a rationale for his study, however, Hanley proposed a fundamental change in the logic of item difficulty. After expressing the oft heard opinion that the concept of item difficulty appears to have little relevance to personality inventories, he redefines it informally by saying (Hanley, 1962, p. 577),

Yet anyone who has responded to a questionnaire knows how much harder some items are to answer than others. The difficult question in the inventory may stimulate reticence, seem ambiguous, or ask for something hard to estimate in oneself.

Because the well-established usage of the concept is potentially as helpful in personality measurement as it has been historically in ability testing, it is necessary to take issue with Hanley's approach.

The lack of precision in Hanley's definition is of small moment, for the concept is being introduced to a new area. What is objectionable is that his definition focuses on response selection or decision making *without regard to the veridical properties of the response*

selected. In justifying this disregard, Hanley has recourse to the cliché that inventory items have no universally correct answers, suggesting that for this reason the assessment of item difficulty requires a method different from the one commonly used in ability testing. By analogy with the measurement of judgment difficulty in psychophysics, he then proposes response latency as an appropriate index of item difficulty in personality measurement.

What Hanley has overlooked here is that the "rightness" and "wrongness" of answers are so basic to the concept of item difficulty that to discard them as inappropriate to personality inventory items is in effect to discard the concept as well. His proposed substitute gets at a kind of difficulty that inheres in items to be sure, but one that can more suitably be designated by a term like *item ambiguity*, *item clarity*, *response equivalence*, or *closeness of response alternatives*. This kind of difficulty no doubt exists in ability tests as well as in personality inventories, but in the former it would certainly be viewed as a source of error and not confused with the kind of stimulus differences which alone make scaling possible (Torgerson, 1958).

To use the concept of item difficulty appropriately in personality measurement requires a framework within which it is meaningful to view responses as right or wrong. Obviously this cannot be done in any absolute sense. But it can be done in a sense that is both universally applicable and perfectly analogous to the correctness of answers on ability items. Just as the right answer to an ability item is the response given by those of sufficient ability, so the right answer to a personality item may be defined as the response given by those sufficiently characterized by the trait being measured. Thus a difficult personality item is one which differentiates those markedly characterized by the trait in question. Their response is the right one, and it occurs with relatively low frequency. An easy personality item, on the other hand, differentiates only those conspicuously lacking in the trait; they alone give the wrong answer.

Several scaling models provide frameworks within which right and wrong answers may be viewed in this fashion. Any model for unidimensional scales composed of monotone items, such as that proposed by Guttman (1950), is appropriate for personality measurement, although those making provision for random variation are undoubtedly the most suitable. For items fulfilling the requirements

of such models, difficulty level may be defined as it generally is in ability testing, i.e. in terms of the relative frequency of the correct response (Nunnally, 1959).

In test construction, test evaluation, or any other situation where there may be doubt that the items fit the model, an index of item difficulty based solely on frequency will of course be inadequate. For whatever they indicate concerning difficulty, frequencies by themselves provide no evidence that items belong on the same dimension. Where the question of dimensionality is unsettled (as it usually is in personality measurement), an index of item difficulty can be made suitable if it takes into account not only the frequency with which the correct or scored response is made, but also *the scale positions attributed to the Ss making the correct response*. One simple but appropriate index, for example, is the mean scale score of those passing each item. Thus,

$$d_k = \frac{\sum s_{jk}}{n}$$

where d_k designates the difficulty of item k , s_{jk} is the scale score (in any unit of measurement) of any subject j who passes item k , and n is the number of Ss so passing. This index can be used for several important purposes in personality measurement, four of which are briefly indicated below.

1. *To evaluate individual items.* (a) The numerical value of the suggested index for any given item will vary from sample to sample, but in no case can any item appropriately show an index value *smaller* than the sample mean. Items with smaller indices discriminate negatively; they should either be reversed in scoring or culled. (b) Also, while numerical values will change, the *relative* difficulty levels of items will tend to remain constant; items showing gross shifts in relative position from sample to sample should be culled as unreliable. (c) Further, within a single sample there should be high correspondence between item frequency and item difficulty, that is, items with similar marginals should also show similar levels of difficulty. If items are ranked both as to frequency of correct response and as to difficulty level (one from large to small), those items whose ranks are markedly discrepant contribute excessive extra-dimensional variance and should be culled.

2. *To test the dimensionality of personality scales.* The relation-

ship just cited provides a convenient basis for testing dimensionality. The correlation between item frequency and item difficulty should approach 1.00 for a unidimensional scale, even allowing for random variation and for the fact that not all item trace lines will have identical slopes.

3. *To help detect invalid protocols.* When the difficulty levels of the items on a personality scale are known, inconsistent performances can be readily identified. In particular, Ss who show correct responses to difficult items but not to easy items may be suspected of faking, of being confused as to instructions, of not belonging to the population for whom the scale was devised, or of some other invalidating factor.

4. *To provide descriptive data concerning the trait or population being measured.* The primary function of the items on a scale is of course to establish scale positions or scores for subjects. When the difficulty levels of the items are explicit, however, item content becomes more useful for conceptual and theoretical purposes.

It should be emphasized that the above uses do not depend on a denial of Hanley's major premise. There is no reasonable alternative to the view that inventory items have no universally correct answers. The limitations of Hanley's view stem from his tacit assumption that universal correctness is logically necessary to a definition of item difficulty based on right and wrong answers. The essential condition for such a definition lies elsewhere, namely, in the assurance of unidimensionality. If a scale is unidimensional, the difficulty level of its items can be established readily. Without the assurance of unidimensionality, on the other hand, indices of item difficulty can represent nothing more than hypotheses to be tested. While unidimensional personality scales have not been common, there is ample evidence that they can be produced (Lumsden, 1961). Further, with the improvement of scaling methods, such as Cattell's (1957) modifications of factor analysis, or the recent development of contextual analysis (Dempsey, 1964), the frequency of such scales is likely to increase.

In short, Hanley's formulation appears excessively limited in two important respects. Not only does it help to perpetuate a common misconception, but also it preempts a central concept to designate a minor error. A view of item difficulty that is closely analogous to its long-established meaning in ability testing is highly appropriate to

personality measurement wherever there is serious concern with dimensionality or with items having monotonic trace lines.

REFERENCES

- Cattell, R. B. *Personality and Motivation Structure and Measurement*. New York: World Book, 1957.
- Dempsey, Paul. "A Unidimensional Depression Scale for the MMPI." *Journal of Consulting Psychology*, XXVIII, (1964) 364-370.
- Guttman, L. "The Basis for Scalogram Analysis." In S. A. Stouffer (Ed.), *Measurement and Prediction*. Princeton: Princeton University Press, 1950.
- Hanley, C. "The Difficulty of a Personality Inventory Item." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 557-584.
- Lumsden, J. "The Construction of Unidimensional Tests." *Psychological Bulletin*, LVIII (1961), 122-131.
- Nunnally, J. C., Jr. *Tests and Measurements*. New York: McGraw-Hill, 1959.
- Torgerson, W. S. *Theory and Methods of Scaling*. New York: Wiley, 1958.

ALTERNATIONS AFTER FORCED CHOICES AS A FUNCTION OF DOMINANCE IN WOMEN

ROBERT D. TARTE¹

University of Michigan

AND

HENRY E. KLUGH

Alma College

DEMBER and Fowler (1958) summarized the work done on spontaneous alternation, a phenomenon that has played an important part in studies of theoretical significance for learning and motivation theory. Since their review Denny² reported the development of a technique to investigate this phenomenon in humans. The method involves presenting *S* with paper and pencil L and T maze outlines. The *Ss* were instructed to start at a dot in the beginning of the maze stem and do anything they wanted to do with the pencil as long as they did not cross the maze outline or lift their pencil from the paper. Denny observed sex differences in alternation behavior with men alternating more frequently than women. He has suggested the possibility of dominance as a personality variable which may be related to this difference. The primary purpose of the present study is to determine if a relationship exists between a measure of dominance in women and tendency to alternate on free choice T's after having experienced forced choice L's. Two subsidiary questions are raised: first, is there an increasing tendency to alternate within dominant or submissive groups following increasing numbers of forced choices; and second, is there a decreasing tendency to alternate within dominant or submissive groups following increasing numbers of free choices, permitted after these forced choices.

¹ The senior author was at Alma College when this research was conducted.
² Denny, M. R., personal communication, 1959.

Method

Subjects. The *Ss* for the study were drawn from 180 female college students who volunteered to participate in the study.

Apparatus. Paper and pencil T mazes and right and left hand L mazes were used. The L mazes were $2 \times 2 \times \frac{1}{4}$ inches. The T mazes were $3\frac{3}{4} \times 2 \times \frac{1}{4}$ inches. Both types were drawn in outline, one to an $8\frac{1}{2} \times 5\frac{1}{2}$ inch sheet of otherwise plain paper. Maslow's Social Personality Inventory (1940) provided a measure of dominance.

Procedure. The Social Personality Inventory was administered to all *Ss* and scored according to the instructions in the manual of directions. The top and bottom 11 percent, 20 dominant and 20 submissive women, were utilized in the remainder of the study. Paper and pencil mazes were administered to all *Ss* individually. Instructions to *S* followed the Denny procedure, "start here," *E* points to dot at the base of the maze stem, "and do anything you want to with the pencil as long as you don't cross the lines or lift your pencil from the paper." When *S* seemed finished *E* inquired, "done?" and on *S*'s assent removed the maze outline and presented the next outline in the series. The inter-trial interval was not more than two seconds on any trial. *S* was given a series consisting of 8 L mazes, one T maze; 12 L mazes, one T maze; 16 L mazes, one T maze, and 4 T mazes in that order and at one sitting. Half of each handedness group received right hand L's and half left hand L's. Alternation on a T was defined as a pencil mark which entered the arm opposite that which was presented on L's for a given *S*.

Results

Scores on the Social Personality Inventory for dominant women ranged from +3 to +157; the median was +39 which falls in the second decile of norm data reported by Maslow. Scores for submissive women ranged from -70 to -113 with a median of -89 which falls in the tenth decile of Maslow's reported norms.

There were seven T's on which comparisons between the 20 dominant and 20 submissive women could be made. On the first T, after 8 forced L's, 14 dominant women alternated, 8 submissive women alternated. On the second T, after 12 additional L's, 16 dominant women alternated as compared with 10 submissive women. On the

third T, after 16 additional L's, 14 dominant women alternated as compared with 10 submissive women. On the remaining 4 T's given to each group, the number of dominant women alternating on each T was 13, 13, 15, and 13. The corresponding frequencies for submissive women were 8, 8, 7, and 8.

For statistical treatment an alternation score was derived for each S. This score was the number of alternations over the first three T's, that is after 8, 12, and 16 forced choices. A comparison of these scores between dominant and submissive groups was significant at the .05 level of confidence (Mann-Whitney U test, two tailed).

A similar score for each S was derived for the last four T's. These were the T's with no interspersed L's. Dominant and submissive women were also significantly different on these trials (.02, Mann-Whitney U, two tailed).

As an additional statistical test, numbers of Ss alternating on both T's one and two, versus those alternating on neither one nor two, for each dominance grouping was subjected to a Fisher's Exact test. The probability of the obtained results was less than .025. A similar test for T's one, two, and three yielded a probability of less than .05.

The analysis supports the first hypothesis that dominant women are more likely to alternate after forced choices than are submissive women, and that this difference continues over four successive free choices.

The last two hypotheses are concerned with within group phenomena. Is there a successive increase in tendency to alternate as a function of increasing the number of forced choices from 8 to 16? The question of significant within group changes over the first three T's after 8, 12, and 16 forced choices was analyzed for each group by means of Cochran's Q. No significant changes in frequency of alternation were noted for either dominant or submissive groups. The final hypothesis concerned the frequency of alternations on each of the final four choices. Proportions of alternation on each of these trials was again analyzed by Cochran's Q. No significant changes in proportions within either group on the final four T's was noted. However, it should be stressed that these free choices may have been influenced by the fact that there were prior forced choices.

Although dominant and submissive women differed in frequency of alternations, there were no changes in frequency of alternations

within each group following increased forced choices or following increased free choices, within the limits of the study. This led to the hypothesis that the observed differences between dominant and submissive Ss might have been due to a sampling error and an additional group of Ss was used to check this hypothesis. These were 30 female college students, primarily freshmen and sophomores, who volunteered to participate in the study. The procedure was similar to the previous study except that only one forced choice L preceded the free choice T. Of 30 Ss, 15 alternated on the free choice and 15 did not. The mean dominance scores for alternators was -32; the mean score for non-alternators was -56. The difference was in the predicted direction. The "t" was 1.94 which is significant with a one tailed test.

We therefore conclude that alternation after forced choice on a paper and pencil maze is a function of dominance in women, but within the limits of this study there is no evidence that it increases with number of forced choices nor declines with number of free choices.

Summary

The purpose of this study has been to determine whether or not a relationship exists between a measure of dominance in women (Maslow's Social Personality Inventory) and tendency to alternate on free choice T mazes after having experienced forced choice L mazes. These mazes were of the paper and pencil Denny variety. The Ss were 180 college women. The 20 most dominant and 20 most submissive Ss were given the mazes which were in the following order: 8 L's, 1 T; 12 more L's, 1 T; 16 more L's, 1 T and 4 additional T's. The results clearly indicate that dominant women are more likely to alternate after forced choices than are submissive women and that this difference is maintained over 4 successive free choices. Thirty additional Ss were run to rule out the possibility of a sampling error. The main effect was reproduced.

REFERENCES

- Dember, W. N. and Fowler, H. "Spontaneous Alternation Behavior." *Psychological Bulletin*, LV (1958), 412-429.
 Maslow, A. H. "A Test for Dominance-Feeling (Self-Esteem) in College Women." *Journal of Social Psychology*, XII (1940), 255-270.

A FACTOR ANALYTIC STUDY OF THE ABILITY TO SPELL

DORIS ALLEN

AND

JOEL AGER

Wayne State University

THURSTONE's statement (1948) that the "ability to spell seems to be quite independent of most other abilities" seems inconsistent with findings that spelling is correlated with numerous measures: visual perception, auditory acuity and discrimination, pronunciation, attitudes toward spelling, interests, habits, general intelligence, vocabulary, age, and education. Reviews of these studies have been made by Horn (1960), Spache (1941a, 1941b), and Williamson (1933). Other studies have shown that personality or temperament is related to spelling achievement (D'Heurle, Mellinger and Haggard, 1959; Kiefer and Sangren, 1925; Schonell, 1936). Guiler (1929), Moore (1937), and Northby (1936) found that the measure of spelling achievement is dependent upon the format or style of spelling test used.

Support of Thurstone's view was found in a factor analysis of spelling in which a "general spelling ability" factor was identified (Knoell and Harris, 1952). However, weaknesses in this study prevent its findings from being conclusive. All 20 measures which were analyzed were obtained from only two tests which leads to the possibility of spurious factors (Guilford, 1952). Furthermore, since only spelling measures were employed, no information was available as to the relationship of spelling to other abilities.

The present study was designed with two purposes: to investigate the relationships between spelling ability and other kinds of variables, and to determine whether spelling tests of different formats varied in their factorial composition.

Procedure

A battery of 11 tests was assembled to investigate this problem, from which 26 variables were obtained. Four of the tests, yielding six measures, were obtained from the regular school evaluation program.

Four spelling tests were used to evaluate the effect of format. Three of these were constructed for this study: a check-list approach (Recognition),¹ a dictation-in-context approach (Oral), and a write-in correction approach. Two scores were obtained from the latter: a recognition score (Write-In: Recog.) indicating the number of items recognized as being correct or incorrect as presented, and a spelling score (Write-In: Spelling) showing the number of words correctly spelled after identified as incorrect. The fourth spelling measure used was the Wellesley Spelling Scale, Form I (Multiple Choice), a multiple choice test with four alternatives per item. All of these tests were untimed.

Spelling items also contributed to two other tests used in this analysis. Aptitude Tests for Occupations: Clerical Routine Aptitude, Form A (Clerical) is a timed multiple choice test which includes equal amounts of name-checking, number-checking, and alphabetizing, as well as spelling, items. One of the school evaluation tests, Sequential Tests for Educational Progress: Writing (STEP: Writing), measures mechanics of writing through the use of multiple choice items and includes some spelling. Both of these are timed tests.

Since it was suspected that spatial ability may be involved in performance on recognition-type spelling tests, the Survey of Space Relations Ability, Form A (Space) was included in the test battery. This is a timed multiple choice test in which the parts contributing to a design must be identified.

An unpublished authoritarianism measure, the Stereopathy-Acquiescence Schedule,² Form I (S-A), was included to determine whether this personality dimension contributes to spelling performance as suggested by D'Heurle, *et al.* (1959). This untimed measure presents 100 ideological statements to each of which the examinee

¹ The terms in parentheses indicate the manner in which the tests are identified in subsequent discussion.

² Use of this measure is through the permission of Dr. George Stern, Syracuse University, to whom the authors wish to express their gratitude.

indicates the extent of his agreement using a 6-point Likert-type scale ranging from "strongly agree" to "strongly disagree." Thirteen scores were obtained from this instrument: ten independent scores (S-A: 1-10) corresponding to ten separate scales of the test, a Stereopathy score (S-A: Stereo.) which is the sum of the first four scales, a Non-Stereopathy score (S-A: Non-Stereo.) summing the remaining six scales, and a Total score (S-A: Total) summing these two.

From the school records, the School and College Ability Test (SCAT) was selected as a general intelligence measure. Three scores were included: Verbal (SCAT: Verbal) based on synonyms, antonyms, and sentence completion items; Quantitative (SCAT: Quant.); and Total (SCAT: Total) representing the sum of these two.

Two paragraph comprehension measures were also taken from the school records: STEP: Reading and STEP: Listening. These differ in that the paragraphs are presented visually for the former and orally for the latter test.

Thus, the 26 variables consisted of 5 measures solely concerned with spelling, 2 containing some spelling along with either perceptual tasks (Clerical) or grammar and punctuation (STEP: Writing), 1 of spatial ability, 13 measures of authoritarianism, 3 of general intelligence, and 2 of paragraph comprehension.

The sample consisted of 100 twelfth grade volunteers, 44 boys and 56 girls, from Denby High School in Detroit. The group was considered to be relatively homogeneous as to age and socio-economic background.

A principal axes solution was obtained from the product-moment correlations of these variables. Factoring was arbitrarily stopped when 95 percent of the common variance had been extracted. A normalized varimax rotation (Kaiser, 1959) was then made of the resulting seven factors. The entire analysis was performed on an IBM 7040 computer.

Results

Seven orthogonal factors were obtained from the analysis of the 26 variables. Loadings greater than .30 were designated significant and were used to interpret the factors. In the following discussion, the significant loadings will be presented in descending order

for each factor.³ Variables which have loadings above the selected critical value on factors other than the one being discussed will have this information presented in parentheses.

Factor 1—General Spelling Ability

Write-In: Spelling	.91
Write-In: Recog.	.89
Recognition	.88
Multiple Choice	.81
Oral	.80
Clerical	.59 (Factor 6 .42)
STEP: Writing	.47 (Factor 3 .59)

This factor is clearly a spelling factor since every variable which loads significantly on it is a spelling measure in whole or in part. None of the non-spelling variables have significant loadings here. In addition, those measures dealing solely with spelling (the two Write-In scores, Recognition, Multiple Choice, and Oral) do not load appreciably on any other factor. The highest loading of any of these variables on another factor was .22 for Multiple Choice on Factor 3; 23 of these loadings on other factors were less than $\pm .10$. This factor accounted for 26 percent of the common variance.

The emergence of a specific spelling factor is consistent with Thurstone (1948) and with Knoell and Harris (1952), as discussed previously. The finding that none of the "pure" spelling tests loaded on other factors suggests that the various formats used do not vary in factorial composition; they may be considered equivalent measures of spelling ability.

Factor 2—"Stereopathy"

S-A: Stereo.	.92
S-A: 2	.86
S-A: 1	.80
S-A: Total	.71 (Factor 4 .67)
S-A: 3	.56 (Factor 5 .47)
S-A: 4	.52 (Factor 5 .44)
S-A: 5	.50 (Factor 4 .33)
STEP: Listening	-.33 (Factor 3 .64)

This factor was termed "Stereopathy" since it corresponds well with the scoring procedure used to obtain the Stereopathy score of the S-A Schedule. As mentioned previously, this score is the sum of the first four scales and also contributes to the S-A: Total score. A companion factor derived from this instrument is factor

³ Hectograph copies of the complete varimax matrix are available from the senior author.

4 which is identified as "Non-Stereopathy," corresponding to its scoring procedure. A similar factor structure was obtained by Lane⁴ in an analysis of this form of the S-A Schedule and Form P (Personality) of the same.

Since this factor represents the contribution of a research instrument and is not involved apparently with spelling ability, no further discussion of either the meaning or the implications of it will be made here.

Factor 3—Verbal Reasoning

SCAT: Total	.90	
SCAT: Verbal	.81	
SCAT: Quant.	.75	(Factor 7 .31)
STEP: Reading	.69	(Factor 7 -.36)
STEP: Listening	.64	(Factor 2 -.33)
STEP: Writing	.59	(Factor 1 .47)
Space	.31	(Factor 6 .45)

The identification of this factor as "verbal reasoning" seems fairly consistent with the nature of the variables found here. However, SCAT: Quant. and Space require further comment. It may be that this factor is confounded with "general reasoning" since tests were not included in this study to differentiate these. On the other hand, SCAT: Quant. may load on this factor due to the verbal presentation of some of the items. The loading of Space may indicate some implicit verbalization employed by the more successful subjects. Perhaps those with higher verbal reasoning skills can verbalize the figural problem more adequately to themselves, and, thus, solve it with greater accuracy and speed than those with lower verbal skills.

With the risk of being redundant, it is interesting to note that none of the spelling tests load on this factor, emphasizing that these kinds of verbal skills are not a factor in spelling performance.

Factor 4—"Non-Stereopathy"

S-A: Non-Stereo.	.94	
S-A: Total	.67	(Factor 2 .71)
S-A: 10	.63	
S-A: 7	.61	(Factor 5 -.34)
S-A: 9	.53	
S-A: 6	.49	(Factor 5 -.35, Factor 6 .31)
S-A: 8	.49	(Factor 7 -.38)
S-A: 5	.33	(Factor 2 .50)

⁴ Personal communication, Dr. George Stern, Syracuse University.

Since this factor corresponds precisely, in its variables, to those parts of the S-A Schedule which contribute to the Non-Stereopathy score, it was so designated.

Factor 5—Unidentified

S-A: 3	.47 (Factor 2 .56)
S-A: 4	.44 (Factor 2 .52)
S-A: 6	-.35 (Factor 4 .49, Factor 6 .31)
S-A: 7	-.34 (Factor 4 .61)

It was not possible to detect any common element among the variables which loaded here since little is known about the exact nature of the scales of the S-A Schedule. Since, however, these variables are all from this personality measure, and, since they suggest some bipolar factor, it may be that some sort of response set is emerging here. This factor accounted for only 5 percent of the common variance.

Factor 6—Perceptual Speed

Space	.45 (Factor 3 .31)
Clerical	.42 (Factor 1 .59)
S-A: 6	.31 (Factor 4 .49, Factor 5 -.35)

Although few variables load on this factor, it seems safe to infer that it represents perceptual speed since the two largest loadings (Space and Clerical) are both for highly speeded tests involving a perceptual task. Although there were other timed tests in the battery, these two were the only highly speeded measures. This factor accounted for only 4 percent of the common variance.

Factor 7—Unidentified

S-A: 8	-.38 (Factor 4 .49)
STEP: Reading	-.36 (Factor 3 .69)
SCAT: Quant.	.31 (Factor 3 .75)

This factor was not identifiable since no common element could be discerned among the three variables which loaded on it. This factor also accounted for 4 percent of the common variance.

Discussion

The general purpose of this study was to investigate the domain of spelling ability and its relationships to other areas of cognitive functioning. The clear factor picture obtained indicates that spelling ability may be an independent skill as suggested by Thurstone (1948) and as found by Knoell and Harris (1952). This conclusion is, of course, restricted to the kinds of variables studied here. It is

not meant to imply that spelling is not related to other kinds of variables.

This study has accomplished the more specific purpose of eliminating several variables as being critical to spelling ability. Spatial ability, as measured here, was not relevant to spelling. Authoritarianism also was not related to it, nor was verbal reasoning. Furthermore, the results of this study indicate that the ability to spell is not influenced by the method of testing. This is suggested both by the high intercorrelations between the various spelling measures, and by the fact that the tests did not differ in their factorial composition.

This last point is in disagreement with those studies mentioned previously (Guiler, 1929; Moore, 1937; Northby, 1936). Several explanations for this are possible. First, the experimental design used by both Moore and Northby confounded order effects with main effects. The same set of words was used for each method of testing spelling; the order of administration was not varied. Guiler pointed out that the order of administration is a significant variable in spelling scores from different tests. The present study differed in that independent lists of words were used. Another explanation is that the skills used in spelling may change with age so that relationships obtained at one age level may not be found at another. The other studies used students in the lower grades, while this investigation used twelfth grade students.

One important implication of this study is that, if spelling is indeed an independent skill, it should receive specific attention and instruction in the school curriculum. In other words, transfer effects should not be expected. Some evidence supporting this comes from other sources. Tyler (1939) found that it is primarily good spellers who show improvement in spelling through secondary learning. Gilbert and Gilbert (1944) found that only good spellers are able to improve their spelling through reading. Thus, it would seem that all but the good spellers would profit from increased time and emphasis devoted to learning to spell.

Summary

A factor analytic study of spelling ability was done to investigate the relationship between spelling and other areas of cognitive functioning such as personality, spatial, clerical (perceptual), and in-

tellectual factors. Another purpose of the study was to determine whether four different methods of testing spelling varied in their factorial composition. The subjects were 100 twelfth grade students. Twenty-six variables from 11 tests were analyzed. Seven orthogonal factors were obtained using a normalized varimax procedure.

The results indicate that all of the common spelling variance is explained by the first factor, identified as "general spelling ability." All of the spelling variables had high loadings on this factor and none loaded on any of the other factors. This suggests that spelling ability is a relatively independent skill. Also, it was found that the four different methods of measuring spelling achievement did not vary in their factorial composition.

REFERENCES

- D'Heurle, Adma, Mellinger, Jeanne, and Haggard, E. "Personality, Intellectual, and Achievement Patterns in Gifted Children." *Psychological Monographs*, LXXIII (1959), No. 483.
- Gilbert, L. and Gilbert, Doris. "The Improvement of Spelling through Reading." *Journal of Educational Research*, XXXVII (1944), 458-463.
- Guiler, W. S. "Validation of Methods of Testing Spelling." *Journal of Educational Research*, XX (1929), 181-189.
- Guilford, J. P. "When Not to Factor Analyze." *Psychological Bulletin*, IL (1952), 26-37.
- Horn, E. "Spelling." In C. W. Harris (Ed.), *Encyclopedia of Educational Research* (3rd ed.). New York: Macmillan Company, 1960, pp. 1337-1354.
- Kaiser, H. F. "Computer Program for Varimax Rotation in Factor Analysis." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIX (1959), 413-420.
- Kiefer, Frieda and Sangren, P. "An Experimental Investigation of the Causes of Poor Spelling among University Students, with Suggestions for Improvement." *Journal of Educational Psychology*, XVI (1925), 38-47.
- Knoell, Dorothy and Harris, C. W. "A Factor Analysis of Spelling Ability." *Journal of Educational Research*, XLVI (1952), 95-111.
- Moore, J. E. "A Comparison of Four Types of Spelling Tests for Diagnostic Purposes." *Journal of Experimental Education*, VI (1937), 24-28.
- Northby, A. S. "A Comparison of 5 Types of Spelling Tests for Diagnostic Purposes." *Journal of Educational Research*, XXIX (1936), 339-346.
- Schonell, F. J. "Ability and Disability in Spelling amongst Educated Adults." *British Journal of Educational Psychology*, VI (1936), 123-146.

- Spache, G. "Spelling Disability Correlates: I. Factors Probably Causal in Spelling Disability." *Journal of Educational Research*, XXXIV (1941), 561-586. (a)
- Spache, G. "Spelling Disability Correlates: II. Factors That May Be Related to Spelling Disability." *Journal of Educational Research*, XXXV (1941), 119-137. (b)
- Thurstone, L. L. "Psychological Implications of Factor Analysis." *American Psychologist*, III (1948), 402-408.
- Tyler, I. K. *Spelling as a Secondary Learning*. New York: Teachers College, Columbia University, 1939.
- Williamson, E. G. "Mental Abilities Related to Learning and Spelling." *Psychological Bulletin*, XXX (1933), 743-751.



A RATIONALE FOR MEASUREMENT IN THE VISUAL ARTS¹

R. MURRAY THOMAS

University of California, Santa Barbara

For more than 40 years psychologists and art educators have been developing tests involving drawings or paintings. The ones of these which are available nationally can be divided into two categories, according to what they have attempted to measure: (1) those focusing on general personality characteristics and (2) those focusing on artistic taste and talent.

In the first category are such instruments as: (a) Goodenough's *Draw-A-Man Test* (1926) and Lantz's *Easel Age Scale* (1955) which are used to measure general intelligence in children as well as some aspect of psychological adjustment; (b) Murray's *Thematic Apperception Test* (1943), Buck's *House-Tree-Person Test* (1947), and more than a score of others which attempt to reveal the examinee's idiosyncratic view of himself and the world (Hammer, 1958); and (c) Welsh's *Figure Preference Test* (1951) which attempts to measure a general creativity factor that pervades all aspects of a person's life.

This paper, however, does not concern itself with the foregoing uses of drawings for general personality appraisal or for measurement of general intelligence. Rather, it focuses on the instruments in the second category, such as Graves' *Design Judgment Test* (1948) and Knauber's *Art Ability Tests* (1935) which have been used for such purposes as determining aesthetic sensitivity or "good taste" and predicting students' success in art courses. Although the past 35 years have seen the publication of a variety of these kinds of tests,

¹The author gratefully acknowledges comments on this paper by David Epperson and William B. Michael.

the problem of adequately measuring abilities to judge and to create art remains unsolved. Unless some new tacks are taken, the prospects for better solutions in the future do not appear bright.

The purpose of this paper is to identify some key reasons for the inadequacies of existing art tests and to suggest a rationale for developing measures that should be more adequate for: (1) predicting success in art, (2) clarifying basic differences between one school's or artist's preferences and another's, and (3) charting the changes over a period of time in an individual's art preferences, his skills of analysis, and his art-production abilities.

Shortcomings of Present-Day Measures

Four major shortcomings of art tests of the past have been: (1) they have measured only one, or at most a few, of the complex factors that constitute art ability; (2) they seemingly have been based upon the assumption that there is a single, correct criterion of good art; (3) they have failed to distinguish adequately among abilities to judge, to appreciate, and to create art; and (4) they have failed to account adequately for the creativity component in art ability.

By first discussing each of these shortcomings, we can build a foundation for the rationale that will follow.

Shortcoming 1: The measurement of too few factors. Although an art product is a unitary thing—a Gestalt—the present writer believes that it can be meaningfully analyzed into a variety of components. It is true that many artists and critics reject this point of view. They believe formal attempts to analyze the components of art works or art ability are futile and that a sort of general aesthetic intuition (which they possess and others lack) is the only valid art appraisal device. On the other hand, art critics in one paragraph will dissect a painting from the aspect of design and in subsequent paragraphs comment on its color combinations, the kind of subject matter pictured, the degree of visual realism portrayed, the surface texture, and the like. These several aspects or factors are themselves often analyzable into even more elemental components. For example, design can be viewed from the standpoints of the type and direction of lines, the varieties of shapes and their placement, the arrangement of space, and so forth.

Not only can we abstract the above aspects from the Gestalt, we can also analyze interactions among them to determine how interrelationships contribute to the art work's overall effect. For exam-

ple, on a simple level of analysis we can inspect the relationship of subject matter to the art medium. (Should delicate ferns be portrayed in cast iron?) Or we can evaluate the interaction of color and subject matter. (Is a purple cow admissible in good art?) We can appraise the relationship of color to line-type and line-direction. (Is pale pink well suited to a chaotic variety of line sizes and directions?) Of course, describing how one factor relates to only one other factor does not provide a complete understanding of interrelationships. A complete understanding can result only when the total interactions of all components are analyzed.

If we can grant the foregoing principal assumption—that an art product is a complex structure of a variety of components—then one key shortcoming of such art tests as the *Graves* and *Meier* scales is obvious. These instruments plumb only a very few of the many factors that combine to make up art ability.

The *Graves Design Judgment Test* (1948), which consists of 90 sets of two or three abstract designs, apparently measures a person's ability to judge the extent to which two-dimensional drawings are organized according to eight traditional art principles—unity, dominance, variety, balance, continuity, symmetry, proportion, and rhythm. A second well known instrument, the *Meier Art Judgment Test* (1942b) of 100 items, is similar in form but includes another variable, subject matter. For instance, one pair of drawings portrays a vase, another pair a city street, a third pair Columbus's ships at sea, and a fourth pair some Japanese women.

Neither the *Graves* nor the *Meier* scale touches upon such variables as color or surface texture or media in three dimensions. Furthermore, the *Meier* instrument, which includes both design and subject-matter variables, is not organized in a way that enables us to separate the examinee's reactions to subject matter from his reactions to the underlying design.

It appears that the ineffectiveness of these scales in predicting students' success in art schools (Gutekunst, 1959; Michael, 1960; Wold, 1960) is due at least partly to their treating only a limited variety of aspects of the art process and product. Measures of the future need to cover a greater range of pertinent variables.

Shortcoming 2: The assumption of a single criterion of good art. There are quantities of evidence to support the contention that artists frequently disagree on what is good art. A range of disagreement is also found among critics and within the lay public. Surreal-

ists assail abstractionists and realists. Sculptors of life-like human figures consider the artistic welders of iron pipes to be junkmen. Other, more catholic critics willingly accept realistic, stylized, surrealistic, and non-objective art works as being equally worthy efforts, provided of course they fulfill specified design characteristics or they display novelty of composition.

This fact that good art, unlike good mathematics and good grammar, does not have standardized characteristics is of prime importance to the constructor of art measures. Either he must build a limited-use instrument which appraises art ability from the standpoint of only one school or of one philosophical position (and label such an instrument with an appropriately limited title), or he must build an instrument which will have universal applicability because its standards can be differentially adapted to the philosophies of different schools of art.

The tests of the past seem to have done neither of these things very adequately. In some cases (such as the *Knauber Art Ability Tests*) the scales have not been based upon any clearly stated philosophy and the items have not been standardized on a well described criterion group. In other cases (Meier, Graves) the only items used in the final form of the test have been ones that were answered in a particular manner by a large percent of art teachers or art students. Such a procedure for item selection might be defensible were we to accept the assumption that there is one agreed-upon criterion of good art. But this assumption is hardly supportable when tastes in art differ so much from one era to another and vary so much among people who claim expertness. It is conceivable that the test makers who include only noncontroversial items (that is, those items eliciting much art teacher agreement) are thereby eliminating other items which might well contain, in the eyes of certain artists, fine qualities of creativity.

The history of art has shown that the innovators who are ridiculed by the majority of their well settled contemporaries (including art teachers) are the people who at a later date may be lauded as the giants of their time. It seems unwise to regard a consensus among teachers or students as necessarily representing a single, correct criterion of good art. Measures of art ability and art judgment in the future should be more adaptable to different schools or philosophies.

Shortcoming 3: Lack of distinctions among abilities to appreciate,

to judge, and to produce art. The ability to judge art does not consist of exactly the same components as the ability to create it. If this were not true, all capable critics would also be capable artists and vice versa. However, it is also true that the critic and the artist apparently have certain skills in common, though the nature of these shared abilities has not yet been clearly determined.

The constructor of measuring devices faces a three-fold task in relation to these receptive and productive aspects of art. First, he is obligated to define operationally the nature of art appreciation, art judgment, and art production. Second, he must devise measures of these operationally-defined characteristics. Third, he needs to validate both the characteristics and their measures in appropriate life situations. The authors of the art tests now available have not satisfactorily completed these three steps. Because of this we cannot say exactly what present-day scales probably measure, but we must be satisfied with estimates. The Meier and Graves tests focus on some aspects of art-design judgment but do not show how the measured factors relate to appreciation or production. The Welsh (1949) test seems to measure preference for complex and unconventional figures. The Horn (1953), Knauber (1935), and Lewerenz (1927) scales demand that the examinee produce some art, but the relation of these tests to judgment and appreciation is not clear.

Hence it appears that more complete and systematic theory, accompanied by appropriate measuring instruments and validation, is needed to solve these problems of the interrelationships of appreciation, judgment, and production.

Shortcoming 4: Inadequate treatment of creativity. Of all the art terms that are inadequately defined and inadequately measured, *creativity* probably deserves to head the list. There are several apparent causes for the inadequacy of creativity definitions. Five of the contributing factors which demand the attention of the test maker are: the difficulty of translating visual impressions into words, the differences of opinion about the components of creativity, the varied experiential backgrounds among people who define creativity as novelty, problems of determining the relationship of novelty to other components of art, and problems of standardizing creativity.

(a) *Translating visual impressions into words.* Verbal communication often breaks down because each of the conversants has in mind a different referent or different quality of art work when he

uses the word creativity. Though the discussants utter the same noise, they intend different meanings. This communication problem is further compounded when one or more of the discussants have not really clarified for themselves what it is about one art work that makes them think it is more creative than another.

The most profitable step toward improving communication would be for the conversants to operationalize their definitions. That is, they need samples of art work at hand with which to illustrate this trait or factor. By showing pairs of art products (one of the pair which they think illustrates creativity, the other which they think does not), their communication should improve.

(b) *Disagreement about the components of creativity.* Is creativity a single component that combines with other factors to make up the work of art? Or is creativity itself a combination of various components or facets? The following sample definitions illustrate the fact that there is no universally agreed upon answer to these questions.

McFee (1961, p. 129) writes: "Creativity . . . refers to people's behavior when they do such things as (1) invent a new pattern, form, or idea, (2) rearrange already established objects, patterns, or ideas, and (3) integrate a new or borrowed factor into an already established organization."

MacKinnon (1961, p. 90), discussing the term as applied to other areas as well as to art, states: ". . . creativity has at least three phases or aspects. It involves a response that is novel or at least statistically infrequent; but that is not sufficient. It must be adaptive to reality; it must serve to solve a problem or fit the requirements of a reality situation. And, finally, there must be an evaluation of the original insight, together with a sustaining and developing of it to the full."

May (1959, p. 263) has defined it as the process of bringing something new into birth.

Meier (1942a, p. 149) writes: "The characteristic creative artist engages in activities that are in a large measure original or at least in those that offer some fresh approach."

Jenkins (1958, p. 113) says that ". . . the specific function of artistic creation is to articulate man's vague apprehensions of the particularity of things, and to embody and present these with great clarity and persuasion."

Lowenfeld (1957, p. 59) defines creativity as "the ability to explore and investigate. . . . In child art creative growth manifests itself in the independent and original approach the child shows in his work."

If communication is to improve regarding creativity, it appears that agreement on a more functional definition needs to be made. The most feasible solution would seem to lie in limiting the meaning to the one component that is common to almost everyone's definition: the element of *novelty* or *originality*. Or we could as well use the terms *uniqueness* or *differentness* as synonyms for this limited meaning of creativity. Those people who customarily include other characteristics as well (as does MacKinnon) still may use these additional facets by regarding them as separate factors that interact with the creativity (*novelty*) dimension.

Throughout the rest of this paper, the term *creativity* will be used to mean *novelty* or *originality* and nothing more than that.

(c) *Different from what?* After we agree to define creativity as uniqueness we must determine: "Unique as compared to what?" This becomes a knotty problem when we recognize that everyone has had a different background of experiences. So the work of art that appears novel to one person may not appear novel to another who has already seen many similar works. An elementary school child who has not viewed a wide variety of paintings may create a drawing whose style is quite novel from the viewpoint of his own experiences, but it is not unusual to the eyes of his teacher who has seen many similar ones among the hundreds of pictures her pupils have painted in the past. Hence the child, compared to his own phenomenological field, is a significant innovator. But from the teacher's viewpoint he is not.

Thus in measuring creative ability we must struggle with this question: *What is the standard for determining novelty?*

(d) *Relationship of creativity to other components of an art work.* Early in our discussion we suggested that an art work can be analyzed into a variety of components or dimensions. The dimension of creativity, as defined above, touches upon all of the other dimensions: color use, design characteristics, choice of subject matter, handling of media, and the rest. An artist can express originality in any or all of these dimensions by treating them in novel ways or by combining them in an unusual manner. The question now is:

How much uniqueness in how many of these dimensions can we admit and still have the work considered "good art"? Or, stated in the opposite way, can an artist be too novel? The most creative artist would be the one who broke with tradition in all dimensions, but in doing so he would produce a work which was so different in so many ways that everyone else might well label his work chaos, not art. So the puzzle remains: *What delicate balance of the traditional and the unique is needed for a work to be "good art"?*

(e) *Standardizing the unique.* Finally, the test maker, in striving to produce norms that enable test users to interpret the results meaningfully, is faced with a most challenging problem when he deals with creativity. Since by its definition creativity is deviation from the usual, how do you make norms for the unique?

In sum, it appears that creativity poses for the test maker some of his most difficult problems.

More Adequate Art Measures

The foregoing survey of shortcomings of currently available art measures suggests several characteristics that more adequate new measures should possess.

1. Provision for measuring a greater variety of factors that compose art products and art abilities.
2. More adequate distinctions among the abilities involved in appreciation, criticism, and production of art.
3. Test norms that are adaptable to different schools or philosophies of art.
4. A more adequate treatment of the creativity dimension.

The remainder of this paper contains a proposal for new art measures which, the writer believes, contain the first three characteristics above. Regrettably, the proposal does relatively little to solve the problem of measuring creativity.

The proposal is organized in five parts: (1) types of devices needed, (2) the bases for norms, (3) tests of art preference, (4) tests of art analysis, and (5) tests of art production.

Types of Devices Needed

Table 1 shows the six categories for which measuring devices are desired if we are to appraise past achievement and to predict future success in both the consumer and the producer aspects of art.

TABLE 1
Categories of Measuring Devices

	Consumer Aspects		Producer Aspects
	Art Preference	Art Analysis	Art Production
Measures of Achievement	Battery I	Battery II	Battery III
Measures of Aptitude	Battery IV	Battery V	Battery VI

In Table 1 *consumer aspects* refers to the process of viewing other people's works, or one's own, in contrast to the process of creating art objects. As suggested earlier, there exists considerable disagreement in the art world about the meaning of terms applied to this consumer function—terms like appreciation, judgment, criticism, and analysis. Of these, appreciation is probably the most common and most confusing—confusing because it has meant so many different things to so many people. To some it simply denotes liking or desiring-to-look-at. To others it has nothing to do with liking or disliking. Rather, it means understanding design structure and art technique. For others, appreciation implies knowing the era during which the artist lived and the conditions under which he created his works. To still others it means all these things.

If our measurement of this consumer aspect is to make better sense than it has in the past. It seems best to avoid the omnibus word *appreciation* and to adopt terms which may be assigned more precise meanings. The present writer considers the most useful solution is to divide the consumer aspect into two basic categories, preference and analysis, and to measure them separately.

This division is intended to distinguish a person's liking of an art work (his primarily emotional reaction) from his ability to analyze its structure and cultural significance (his primarily intellectual reaction). In the past, art tests have not permitted this distinction, but it seems important that those of the future do. For instance, it is quite possible for a person to understand the design, color use, and historical significance of a painting, yet not like to spend time in its presence. He understands the work but does not enjoy it. A second person may both understand and enjoy the picture. A third may be quite unable to analyze the work but finds himself enthralled by its appearance. Thus if tests can enable us to distinguish between taste

and analytical talent, the nature of these two major components of the consumer aspect can be better understood.

The term *art production* in Table 1 refers to the skills involved in creating one's own works.

The terms *achievement* and *aptitude* are used here with the same meanings they convey to users of standardized tests. Achievement devices focus on what the individual has acquired from his past experiences. Aptitude devices focus on the probabilities of his future success.

As the table suggests, analytically we can distinguish six types of test batteries needed to carry out the desired art measurements. But in practice it is possible that the same battery, or at least some of the same sub-tests, can serve as measures of both achievement and ability.

The Bases for Norms

Earlier we noted the difficulties involved in trying to create general norms for art tests in light of the fact that there is no agreed upon standard of "good art."

The best solution to the norms problem appears to lie in providing answer keys and norms that adjust themselves to each different philosophy of art or school of art. It can be accomplished in this manner. A scoring key is not established by the test constructor. Rather, any artist or critic who considers himself an authority can take the tests. His answers to the items will then constitute the answer key which is appropriate to his philosophy or school of art. Thereafter, when other subjects take the tests, their art talents are gauged by the degree to which their scores approximate the "master's." We thus may be able to have a Salvador Dali key, a Norman Rockwell key, a Picasso key, and so on. After a large number of people take the batteries, their tests can be rescored in as many ways as there are keys available, and different norms can thus be established for each scoring system.

Such a scheme obviously parallels those found in the area of interest testing. Just as a student's profile on interest inventories indicates the degree to which he is like people in certain vocations, so this art scoring plan enables him to see the degree to which his preferences or analyses or production skills are like those of various artists or critics.

Scoring keys and norms derived in this manner promise benefits for both theorists and test users. For instance, theorists may analyze the patterns of test answers given by different artists to determine more clearly the character of their concepts of art and the extent of agreement among them.

Educational counselors may look forward to better predictive validity for art tests. In the past they could not be sure that the test makers' concept of good art coincided with the standards held by the instructors of the art classes or art schools in which students would later enroll. This factor apparently has been part of the cause of low validity coefficients between art tests and instructors' ratings of student success. But with scoring keys made adjustable to each artist's or art teacher's beliefs, higher coefficients should result.

These tests may also help an art student chart his own evolution. By taking the batteries at various stages of his career and comparing the pattern of his responses from one time to another, he may better understand the changes in his philosophy, preferences, skills of analysis, and production talents.

Test of Art Preference

The following section proposes a new test of art preference. The proposal focuses on the test's (1) purpose, (2) type of item, (3) contents, (4) format and mode of administration, and (5) limitations and persisting problems.

Test purpose. The goal is to identify the criteria a person uses in deciding which art works he likes more than others. The test does not attempt to determine the extent to which these criteria are conscious or subliminal nor the degree to which they are applied intellectually or emotionally.

It would be desirable to have two parallel forms which would permit retesting without the chance of the student recalling specific items. However, either of the two forms could be used for measuring either achievement or aptitude. This suggestion to use the same test for estimating both past learning and future potential is based upon two assumptions: (1) that a person does not need any training to express a preference, so subjects with training and those without have equal chances to make a response to each item, and (2) that art preferences are somewhat stable in spite of a particular type of training; thus a person who on a pretest expresses preferences simi-

lar to those of a given artist or school is more likely to show this similarity after training in that school than is a person who on the pretest expressed quite different tastes.

Type of item. The test items are all of the same type. Each consists of a pair of pictures to which the subject is asked to respond in one of three ways: (1) I prefer the left-hand picture, (2) I prefer the right-hand picture, or (3) I have no preference for one over the other. This is basically the same type of item used in such familiar scales as the Graves, Meier, and Welsh. The only new feature is the addition of the "no preference" category.

Test contents. Earlier we suggested that art products can be analyzed into components. The test items would be designed to measure the subject's reaction to these components or dimensions in an effort to determine in which way they affect his likes and dislikes in art.

For example, one of the components that can be expressed as a scale is labeled a *realistic-nonobjective* dimension. This is a continuum that ranges from photographic realism at one end through different degrees of stylized representation in the middle to abstract and nonobjective art at the other extreme. Some people's art preferences lie at one point on this scale, some at another. Still others accept several or all points as equally desirable. To determine where an individual stands on this dimension, several test items are designed with the pair of pictures differing from each other only in the degree of visual realism they represent. All other factors, such as basic design and color scheme, are held as nearly constant as possible (Figure 1). The pattern of an individual's answers on this series of items should suggest how the realistic-nonobjective factor influences his art preferences. If he selects the "no preference" answer for all items, we seem warranted in concluding that he accepts all degrees along the dimension as equally worthy. (The practice followed by many test makers of forcing a choice between the two pictures would, in this test, do violence to an examinee's honest acceptance of two styles of art as equally pleasing.)

In addition to the realistic-nonobjective dimension, several other components of art works can be expressed as scales, and items can be designed to determine a person's position on each dimension. Such scalable factors include the design principles of dominance-subordination, formal-informal balance, and rhythm, all of which can be expressed in color and in black-and-white.

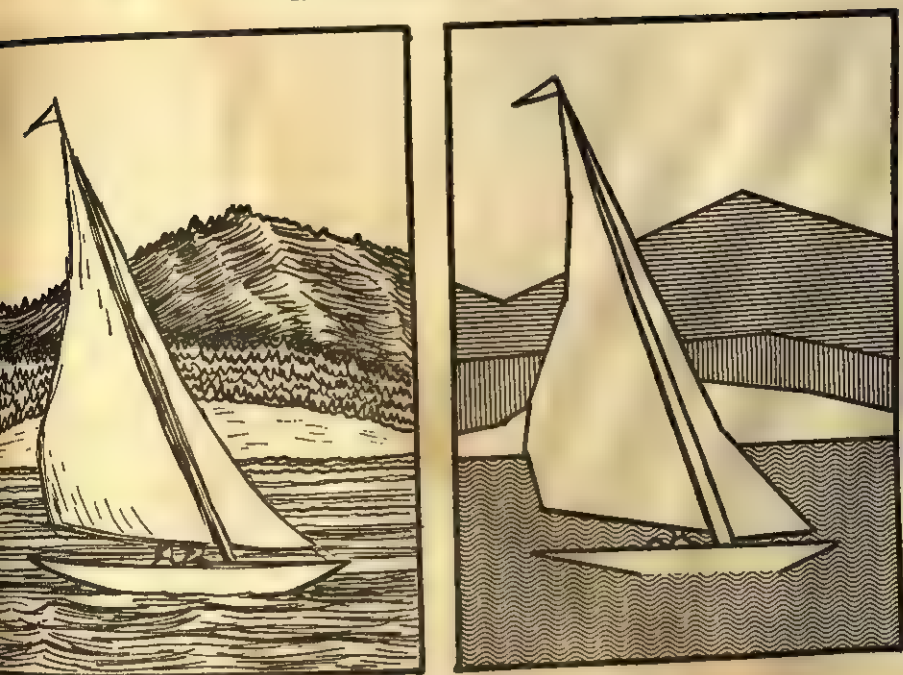


Figure 1

Not all components of art work can be conveniently scaled. Some represent discrete items or classes. This is true of art subject matter. Some people prefer seascapes to still-lives. Others prefer nudes to landscapes. Since these variables do not lend themselves to placement on a continuum, the pairs of pictures which are used to investigate such subject-matter preferences would require choices between categories rather than points on a scale (Figure 2).

The interaction among such components, and the way this interaction affects an individual's preferences, can be estimated by analyses of the relationship of his answers to various combinations of items.

In the foregoing paragraphs a variety of components or factors has been suggested as the basis for the test's contents. The sources of these have been other art tests, descriptions of variables in art textbooks, and the present writer's analysis. These components are proposed as a starting point. After items focusing on them are developed and tried out with a quantity of subjects, the adjustments necessary to improve the categories (combine some, add new ones) can be sug-

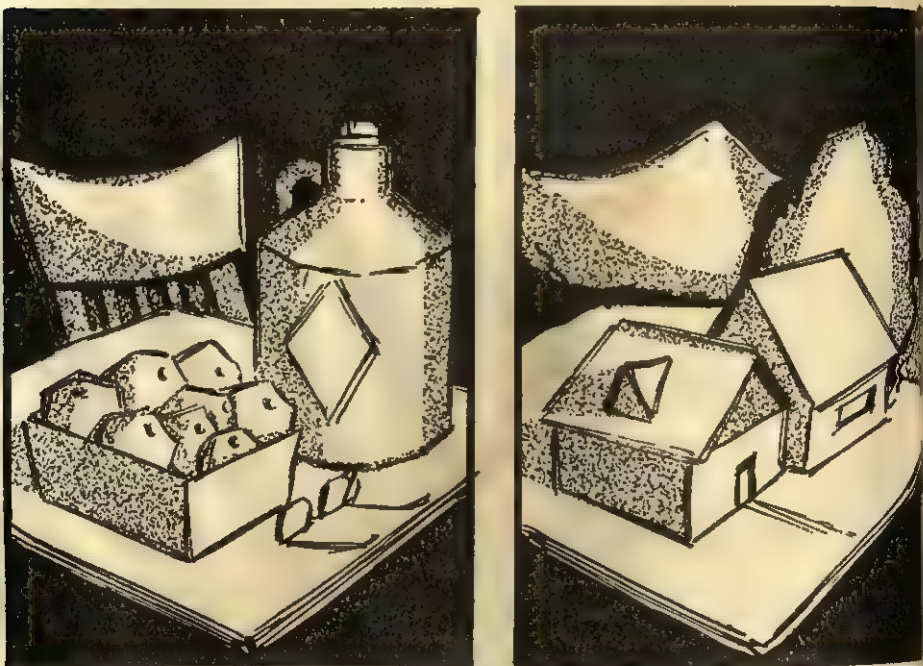


Figure 2

gested by factor analyses and through questioning test subjects about their choices.

Format and administration. Tests used today which require a response to sets of pictures are offered to the examinee in booklet form. But our proposed new test would be in the form of 35 millimeter photographic slides, each displaying a pair of pictures to which the subject responds on an answer sheet. The test is proposed in this form for the following reasons.

1. Currently available tests are mostly limited to black and white plates because color plates are costly. But to test a greater range of art factors, the new test should also include colored reproductions of paintings, drawings, sculpture, architecture, ceramics, and the like. The high costs of printing color plates would prohibit the inclusion of a desired quantity of such productions in an art-test booklet which could appeal only to a limited market. On the other hand, 35mm slides can be produced in smaller quantities at a reasonable cost with promise of good color rendition. A school would need only one set of slides to test any quantity of students.

2. Slides can be used either for large-group, small-group, or individual testing.

3. Revisions of the test could be accomplished readily through the substitution of slides or the rearrangement of their sequence.

The test would be administered in a semi-darkened room. Each slide would be projected for a set interval (probably 20 seconds) during which time the examinees would view the pictures and record their preferences on the answer sheet.

Limitations and persisting problems. Perhaps the main limitation is that the test, at least in its initial stages of development, would probably not plumb all components of art preference. In this regard, the matter of interaction among components likely poses knottier problems than the matter of identifying single components. Factor analyses following trial administrations of the test should help correct this limitation.

In art-preference testing, the creativity component poses another problem. The principal difficulty lies in the fact that, as noted before, each person brings a different background of experience to the pictures in the test. What is unique to one person may not be so to another. If the subjects were asked to indicate which one of two pictures was the more unique or original, we might conclude that we were tapping a creativity reaction. But in this preference test we are measuring only for likes and dislikes, not for analyses. Thus such a question would be inappropriate here, though it can be used in the art-analysis battery. In light of the preferences required here, it seems that it would be difficult to tease out a creativity component.

Tests of Art Analysis

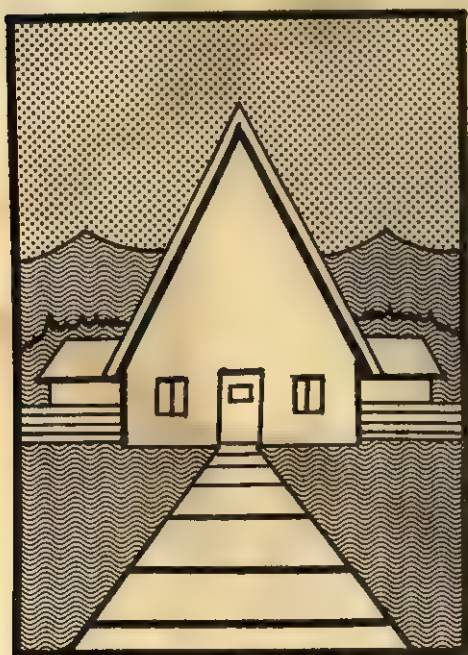
These tests stress the subject's understanding of art rather than his likes and dislikes. They focus on the intellect, not the emotions. Two batteries are recommended. The first assumes little or no art instruction on the part of the examinee; it focuses on his potential for profiting from instruction. The second tests achievement; hence it assumes a quantity of prior training. Though it focuses on the past, it should also prove useful for predicting the individual's potential for analytical work at a higher level.

Tests' purposes. They are designed to determine the individual's potential (aptitude battery) and skill (achievement battery) in: (1) analyzing the design and color structure of art works, (2) iden-

tifying art media and terminology applying to media and techniques, (3) identifying degrees of creativity in art works, and (4) recognizing notable historical developments in art.

Types of items and their content. Because the nature of the aptitude battery is somewhat different from that of the achievement battery, it is appropriate to discuss them separately.

As already noted, the purpose of the aptitude battery is to estimate a person's potential for learning to analyze art. The battery attempts this by simulating a teaching situation that is followed by immediate testing. For example, the subjects are shown a pair of pictures and are told why one is said to possess formal balance and the other informal balance (Figure 3). Then, to determine whether the subjects have learned this distinction from the one exposure, a quartet of pictures is immediately shown and the subjects are to mark on their answer sheets which of the four shows the most formal balance and which the most informal (Figure 4). Next, another pair



1



2

Figure 3. Subjects are instructed: "Picture 1 has formal balance. This means that the left half of the picture is a mirror image of the right half. Picture 2 has informal balance because different lines and shapes are used in the left half to balance those of the right half."

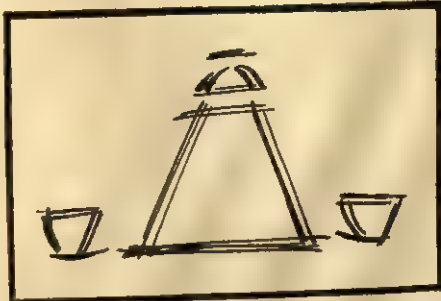
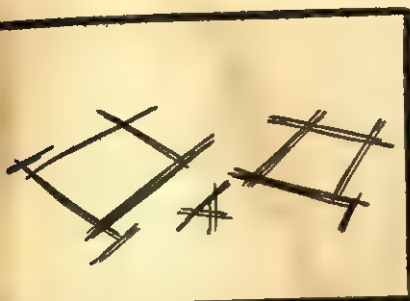


Figure 4. Subjects are instructed: "On line 20 of your answer sheet, write the number of the design that has the most formal balance. On line 21, write the number of the design that has the most informal balance."

of pictures is shown and a second art principle or characteristic is explained. This is followed by a set of pictures whose purpose is to measure the extent of the students' understanding of this second principle. The test continues in this same pattern of teaching-items paired with test-items. In such a manner the battery proposes to estimate which subjects can most readily understand and apply concept of design and color, media and technique, and art history.

The ability to analyze creativity will not be measured in this fashion but will be determined by a series of items, each requiring the subject to tell which picture of a pair is the more unique.

Following these two sections (the teach-and-test items and those on creativity) a final section is included with items to measure for all of the characteristics taught earlier in the battery. The purpose is to determine how well the learning has been retained with the passage of time and in light of the possibility of subsequent items interfering with the remembering of earlier ones.

The foregoing proposed aptitude battery, therefore, is designed to:

(1) estimate the ability to understand art-analysis concepts, (2) determine what creativity (as novelty) means to examinees, and (3) test for short-term memory of art-analysis concepts. Such an aptitude-measuring instrument is built on the assumption that the students who learn and recall the concepts on the test most adequately will also be the ones who will be the finest art analysts and critics following semesters or years of training.

The achievement battery consists of four sub-sections, three of which contain strictly verbal items as well as verbal-pictorial ones.

Design and color analysis. Two kinds of items are recommended. The first consists of four pictures or art plates from which the student selects the one that best answers the question posed on the test booklet. The purpose of these multiple-choice items is to determine the extent to which the subject can apply analytical concepts to art works (Figure 5). The second kind of item is also multiple-choice, but in this case it is entirely verbal. The purpose is to establish the extent to which the examinee understands analytical terms and their verbal definitions. For example:

Line-direction dominance in a drawing means:

- (A) a main line near the center is darker than all others;
- (B) a majority of the lines go in a single direction;
- (C) more lines go in one direction than in any other direction;
- (D) all lines come together at the same point in the background.

Marked discrepancies between a student's success with the pictorial items and the verbal ones would appear to reflect an unevenness between his ability to talk about art and his ability to apply concepts in practical analysis.

Media and technique analysis. Two types of items are also recommended in this section. The first requires the examinee to view a photograph or an art work and to recognize from multiple choices which medium or technique it represents. For instance, the choices from which he selects for a given picture might include: oil paint, transparent water color, line and wash, and pastel crayon. Or he may be presented with four photographs and be expected to indicate which one represents a particular technique.

The second major variety of item does not involve pictures but only verbal multiple-choice items which measure the subject's abil-

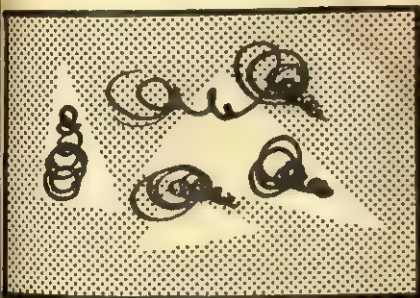
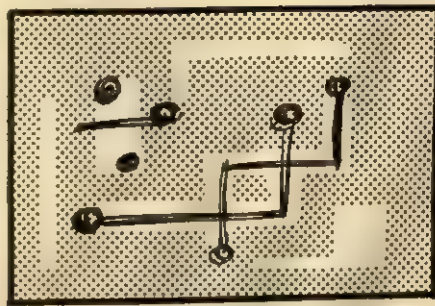
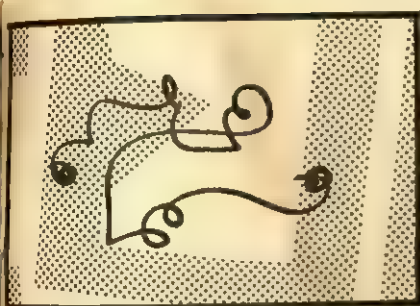


Figure 5. Subjects are instructed: "On line 41 of your answer sheet, write the number of the design that shows the greatest value dominance."

ity to recognize descriptions of art media and techniques and the terminology commonly applied to them. For example:

Gouache refers to painting with:

- (A) paints containing a large amount of oil;
- (B) transparent water colors;
- (C) water colors made nontransparent by adding white;
- (D) oil paints thickened with the addition of beeswax.

As with the design and color items, a marked discrepancy between the student's score on the first and second portions would suggest unevenness between his verbal knowledge and his ability to apply his understanding of media and technique when he views art works.

Creativity analysis. Each item in this section would consist of a pair of pictures. The examinee would be required to indicate which one of the pair is more unique or novel. This item, like its counterpart in the aptitude battery, should reflect how the subject's own experiences and phenomenological field determine for him what is novel. Comparisons between his choices in this section and his

choices in the art-preference test might reveal the extent to which novelty influences his preferences.

Historical analysis. Every item in this part would consist of a picture to which the subject would react by marking a multiple-choice question. Each of the following kinds of art history knowledge would be sampled by several items: (1) subject matter and style of notable painters, (2) subject matter and style that dominated important painting eras, (3) architectural styles and the years of their popularity, (4) sculptural styles during given eras, and (5) styles in other media such as pottery or tapestry of different eras, and nations.

Format and administration. The foregoing series of art-analysis achievement tests should be administered in two parts. One part would consist of 35mm photographic slides projected onto a screen in a semi-darkened room or onto a daylight screen in a moderately light room. The slides would be used for presenting the portions of the design-color, media-technique, creativity, and history tests that require pictures. Each student would mark his answers in a test booklet containing multiple-choice items appropriate to each sequence of slides. The aptitude battery would consist solely of slides.

No oral explanation on the part of the tester would be needed during the administration of items on the achievement battery. However, the tester would need to explain the principle or characteristic being taught in the teach-and-test portion of the aptitude battery.

The second portion of the achievement battery would be administered in the same manner as any typical paper-pencil test because the items are all verbal ones relating to design-color and media-technique analysis.

Limitations and persisting problems. The use of photographic slides poses several problems. Students may sooner become fatigued glancing back and forth from the screen to their answer sheets in a semi-darkened room than they would under the more typical paper-pencil test situation. Providing optimum lighting conditions (enough light to read the answer sheets but not so much that the projected slide is only dimly seen) may be difficult to accomplish in certain rooms.

Another problem is that of ensuring that the items adequately sample all of the areas considered important by artists of varied philosophical bents.

Tests of Art Production

Probably the greatest interest in art tests has stemmed from a desire to predict ability to produce art works rather than to predict preferences or critical abilities. The best known scales requiring the examinee to produce drawings, or to improve sketches containing a fault, are Lewerenz's (1927), Knauber's (1935), and Horn's (1953). However, the available meager evidence of the validity of these measures for predicting success or for judging achievement following training has not warranted confidence in their widespread use (Buros, 1941, pp. 143-144; 1949, pp. 257-258; 1953, p. 223; 1959, pp. 376-377).

The present writer's proposal for tests of production is offered with the hope that it will serve as a better predictor of art production. But this hope is accompanied by a recognition of the shortcomings it shares with presently available measures.

Tests' purposes. The intention is to determine an individual's aptitude for, and present skill in, drawing and modeling. The drawing test is the easier to administer because it requires only a pencil and paper, whereas the modeling test requires clay or similar plastic material. It is recognized that the tests proposed here will not be of equal worth for measuring potential or skill in all varieties of visual arts and crafts, such as pen sketching, oil painting, weaving, sculpturing in wood or stone, jewelry making, and the like. However, it is assumed that the drawing-test items will be of some value in reflecting potential for, and present skill in, painting, printmaking, and similar pursuits which seem to depend upon many of the subskills that constitute drawing ability.

Types of items and their contents. The tasks posed on both the aptitude and the achievement tests are identical. These are merely two forms of the same test.

In both the drawing test and the modeling test there are three sections: (1) perception-coordination, (2) composition, and (3) theme interpretation. All items require the examinee to produce a drawing or create a model from a plastic material.

In the drawing test, the perception-coordination section focuses on habits of perceiving objects and on the hand coordination skills needed for recording these perceptions. The items treat: (A) accuracy in copying forms, (B) skill in sketching essential form of more complex objects, (C) perception of light and shadow patterns, and

(D) perspective. Some items require the individual to add or alter elements in drawings, such as placing the shadows in a landscape or altering incorrect perspective. Others require him to sketch objects projected briefly from photographic slides. Still others offer drawing elements which the examinee is asked to use in solving perspective problems (Figure 6). It is apparent that these items treat basic skills traditionally taught in drawing courses.

The composition section of the drawing test offers several shapes and lines which the examinee is to organize into compositions of several types (Figure 7).

The theme-interpretation portion suggests moods or themes to be represented in drawings. It is expected that creativity will be tapped more readily by the composition and theme items than by those in the perception-coordination section.

In the modeling test the perception-coordination section requires the examinee to use plasticine or water-base clay or a dough material to reproduce an object or figure that is shown from two views in a photograph. The purpose is to judge his eye-hand coordination and command of the plastic medium.

The composition section offers a picture of several three-dimen-

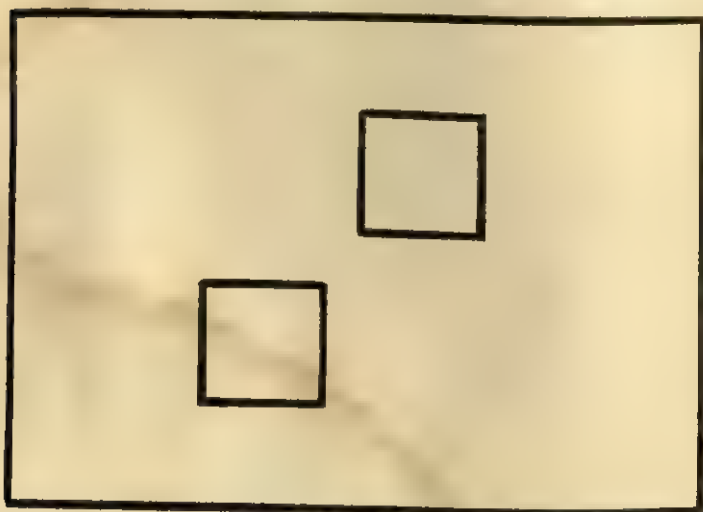


Figure 6. Subjects are instructed: "In the four rectangles marked N on your answer sheet, redraw these two squares in four different manners so that one square appears to be farther in the distance than the other. Add any other lines or shapes you wish and alter the squares in whatever manner seems best to give an idea of distance."

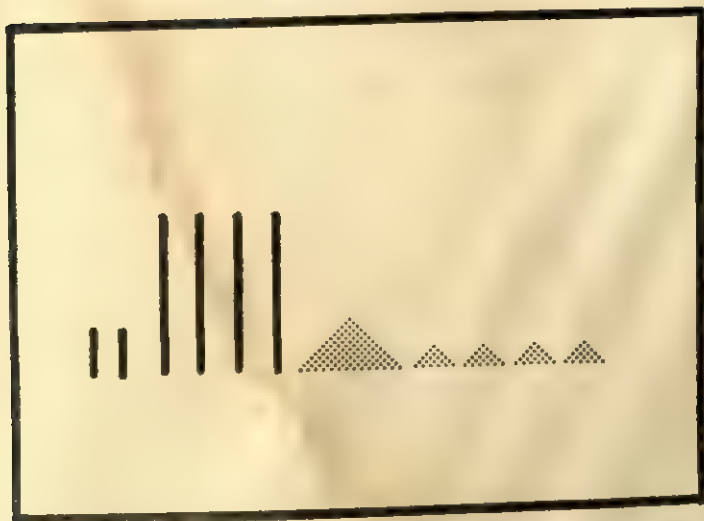


Figure 7. "Using only the lines and shapes shown in this rectangle, draw three different compositions in the spaces on your answer sheet that are marked P, Q, and R. Follow the directions below for creating each composition:

"In rectangle P: Draw the six lines and three triangles in a pattern which you find most interesting and satisfying.

"In rectangle Q: Draw the six lines and three triangles in a pattern which you think best expresses *conflict*.

"In rectangle R: Draw the six lines and three triangles in a pattern which you think best expresses *peace and stability*."

sional forms which the individual is to reproduce in clay and organize into a composition.

The theme-interpretation portion suggests moods or themes which he is to express in model form.

Format and administration. The format of the tests has been implied in the description of items. Part of the items consist of instructions and drawings printed in a test booklet. The others consist of photographs, some in the form of slides (such as shapes to be reproduced in quick sketches) and some in the form of half-tones printed in the booklet (such as figures to be modeled).

The drawing test could feasibly be administered to a larger group than could the modeling test, which requires studio or laboratory tables for the modeling activities.

Scoring procedure. Similar to the scheme used with the Horn test, the criteria which the scorer uses to judge the adequacy of the ex-

aminee's products are in two forms: (1) pictures of products which fulfill the criteria in varied degrees of goodness and (2) verbal descriptions of characteristics which adequate and inadequate products possess.

It is recalled that the proposed tests are to be adaptable to the standards of different artists. Hence the best solutions to the drawing and modeling tasks are not established by the test maker but by the particular artist or critic or school against which the examinee is being judged. Unlike the procedure used with the art-preference and art-analysis scales, the answer key for production tests is not established by having the expert (artist or critic) take the test so his answers can serve as the key. Rather, for each item the expert is furnished pictures and verbal descriptions that represent the range of solutions examinees would produce for the test problems, and he ranks these pictures and descriptions in light of his own ideas of good art. Probably the most feasible ranking system is derived by his assigning each example-picture to one of five numbered categories that represent grades of goodness. Subsequently, when someone else administers the tests, the score assigned to a student's drawing or model will be the number of the picture or description that most closely resembles the student's product.

It also will be desirable to weight different items of the test differently in deriving an overall score. This is because it is unlikely that the artists and critics who determine the various scoring schemes will agree upon which subskills measured by the items are most important in determining art ability. A Norman Rockwell type of artist will probably consider the light-and-shadow and perspective items more essential for art success than will a Picasso type. In order that this difference be reflected in the total score an examinee receives under each of the scoring systems, the criterion artist or critic can estimate the importance to overall art ability which each item represents. This estimate can be reflected in a numerical weighting assigned each item. The most appropriate magnitudes for these weightings probably can be determined only through comparing various weightings with the criterion-expert's overall judgment of a group of examinees' abilities. That is, when a sample group of examinees are ranked by the criterion-expert on art ability or potential, a high correlation should exist between these ranks and the examinees' total test scores. Thus it appears reasonable that the

most appropriate pattern of weighting items is the one that accomplishes this high correlation.

Limitations and persisting problems. A number of questions about possible limitations of the foregoing scheme cannot be answered until the tests are tried out.

Several questions about validity are: Do the tasks required by the test reflect accurately the most important characteristics of art talent from one artist's viewpoint but not from another's? How efficiently will scores on the drawing and modeling tests predict an individual's potential or his current skill in other art pursuits, such as oil and water-color painting, silk-screen printing, etching, and the like?

In regard to scoring, how many different examples of possible criteria-pictures must be provided so the scorer can confidently mark each examinee's products? What does the scorer do about a picture that is so unique and creative that it is similar to none of the criteria-pictures or descriptions?

How many well-known artists and critics will cooperate in establishing scoring schemes to represent their art philosophies? How successfully can a total score on the production tests correlate with the expert's judgments of an individual's overall art ability? In other words, can the whole of an artist's ability be accurately represented by a sum of the weighted parts of the test?

There are also limitations imposed by the administration scheme. Although the drawing test may be administered to a rather large group at one time, the modeling test cannot, because it requires somewhat unusual test conditions. In the drawing test, some items require the use of projected slides—a fact which makes it necessary for students to draw in a room that is not fully lighted.

Conclusion

A rationale has been proposed for measuring art aptitude and achievement by the use of three sets of tests, one focusing on art preferences, another on skills of analysis, and the third on art production talent. Although the proposed instruments do not solve all of the difficulties that plague this area of measurement, the writer believes that they improve on the problems of: (1) measuring the variety of components of art talent, (2) distinguishing more adequately among art preferences, skills of analysis, and production

abilities, and (3) adapting test norms to varied philosophies of art. To a much less satisfactory degree the rationale deals with the problem of measuring creativity (defined as novelty).

Although the new tests, now being constructed, should aid the guidance worker or art teacher who wishes to measure potential and achievement, it is likely that these measures will be of greater use than are other existing ones to researchers who wish to investigate the underlying personality characteristics and experiential background of people who hold different concepts of art.

REFERENCES

- Buck, J. N. *The House-Tree-Person Test*. Colony, Virginia: J. N. Buck, 1947.
- Buros, Oscar K. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: The Mental Measurements Yearbook, 1941.
- Buros, Oscar K. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949.
- Buros, Oscar K. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: The Gryphon Press, 1953.
- Buros, Oscar K. *The Fifth Mental Measurements Yearbook*. Highland Park, N. J.: The Gryphon Press, 1959.
- Educational Testing Service. *Graduate Record Examinations Advanced Fine Arts Test*. Princeton, N. J.: Educational Testing Service, 1951—
- Goodenough, Florence. *Measurement of Intelligence by Drawings*. New York: Harcourt, Brace, and World, 1926.
- Graves, Maitland. *Graves Design Judgment Test*. New York: Psychological Corporation, 1948.
- Gutekunst, Josef Grant. "The Prediction of Art Achievement of Art Education Students by Means of Standardized Tests." Unpublished doctoral dissertation, Temple University, 1959.
- Hammer, Emanuel F. *The Clinical Application of Projective Drawings*. Springfield, Ill.: Charles C Thomas, 1958.
- Horn, Charles C. *Horn Art Aptitude Inventory*. Chicago: C. H. Stoelting, 1953.
- Jenkins, Iredell. *Art and the Human Enterprise*. Cambridge, Mass.: Harvard University Press, 1958.
- Knauber, Alma Jordan. *Knauber Art Ability Test*. Cincinnati, Ohio: A. J. Knauber, 1935.
- Lantz, Beatrice. *Easel Age Scale*. Los Angeles: California Test Bureau, 1955.
- Lewerenz, Alfred S. *Tests in Fundamental Abilities of Visual Art*. Los Angeles: California Test Bureau, 1927.
- Lowenfeld, Viktor. *Creative and Mental Growth*. New York: Macmillan, 1957.
- MacKinnon, Donald W. "Characteristics of the Creative Person:

- Implications for the Teaching-Learning Process." In G. Kerry Smith (ed.), *Current Issues in Higher Education*, 1961. Washington: National Education Association, 1961.
- May, Rollo. "The Nature of Creativity." *ETC*, XVI (1959), 261-288.
- McAdory, Margaret. *McAdory Art Test*. New York: Bureau of Publications, Teachers College, Columbia University, 1929.
- McFee, June King. *Preparation for Art*. San Francisco: Wadsworth Publishing Co., 1961.
- Meier, Norman Charles. *Art in Human Affairs*. New York: McGraw-Hill, 1942. (a)
- Meier, Norman Charles. *Meier Art Tests: I, Art Judgment*. Iowa City: Bureau of Educational Research and Service, State University of Iowa, 1942. (b)
- Michael, William B. "Aptitudes." In Chester W. Harris, *Encyclopedia of Educational Research*. New York: Macmillan, 1960.
- Murray, Henry A. *Thematic Apperception Test*. Cambridge, Mass: Harvard University Press, 1943.
- Varnum, William Harrison. *Selective Art Aptitude Test*. Scranton, Penn.: International Textbook Co., 1946.
- Welsh, George S. *Welsh Figure Preference Test*. Palo Alto, Calif.: Consulting Psychologists Press, 1949.
- Wold, Stanley G. "A Comparison of College Students' Performance on Selected Art Tasks and on the Graves Design Judgment Test." Unpublished doctoral dissertation. University of Minnesota, 1960.



SECONDARY SELECTION IN NAVAL AVIATION TRAINING¹

JAMES R. BERKSHIRE,
ROBERT J. WHERRY, JR.,
AND
RICHARD W. SHOENBERGER
U. S. Naval School of Aviation Medicine

ONE of the most important tasks in an aviation training program is that of deciding whether a student who is in difficulty should be "dropped" or given another chance. Correct decisions in such cases minimize training costs and reduce the numbers of unsatisfactory men who reach operating squadrons. Incorrect decisions mean either wasting aircraft and instructors on men who fail later, or the dropping of men who, if given another chance, would go on to become good aviators and good officers. Modern aviation training is expensive—a man who leaves the training program after he has completed basic flight training may well represent a lost investment of more than a hundred thousand dollars. While the earliest portion of training (pre-flight ground school) costs only about \$100 a week, once the man starts flying the costs mount rapidly, ranging from about \$500 per week in a light primary trainer to about \$5500 per week in operational types of jet aircraft. Since, under current selection standards, about one third of the men who start training do not finish, the total amount of money spent on men who fail to complete is quite large. Consequently, procedures which aid the early identification of men with high failure potential can save millions of dollars.

¹ Opinions or conclusions contained in this report are those of the authors. They are not to be construed as necessarily reflecting the views or the endorsement of the Navy Department.

Portions of this paper were read before the American Psychological Association Meetings, September 1962.

TABLE 1

Scores and Grades Available on Naval Aviation Students

1. Before selection	Aviation Qualification Test
2. " "	Mechanical Comprehension Test
3. " "	Spatial Apperception Test
4. " "	Biographical Inventory
5. " "	Age
6. " "	Years Education
7. Second week	Incoming Mathematics Test
8. Third week	Incoming Jump Reach
9. " "	Incoming Sit-Ups
10. " "	Incoming Speed Agility
11. " "	Incoming Chins
12. " "	Incoming Step Test
13. Sixth week	Mathematics Final Grade
14. Ninth week	Peer Rating
15. " "	Physics Final Grade
16. " "	Trampoline
17. " "	Gymnastics
18. Eleventh week	Navigation Final Grade
19. " "	Study Skills Final Grade
20. Thirteenth week	Outgoing Jump Reach
21. " "	Outgoing Sit-Ups
22. " "	Outgoing Speed Agility
23. " "	Outgoing Chins
24. " "	Outgoing Step Test
25. Fourteenth week	Engineering Final Grade
26. Fifteenth week	Aerodynamics Final Grade
27. " "	Physiology Final Grade
28. " "	Naval Orientation Final Grade
29. " "	Leadership Final Grade
30. After Pre-Solo Stage	Pre-Solo Flight Grade
31. After Precision Stage	Precision Flight Grade
32. After Transition Stage	Transition Flight Grade
33. After T/28 Precision Stage	T/28 Precision Flight Grade
34. After Acrobatics Stage	Acrobatics Flight Grade

Normally a decision about a failing student will be made by taking into consideration available information on the past performance of the individual—his grades; his aptitude test scores; his interests, measured or expressed; his judged motivation, etc. In aviation training such measures accumulate rapidly; Table 1 shows the items of quantitative information that become available on students during their early training. The administrator who must make decisions about marginal students is soon embarrassed by the availability of more performance data than he knows what to do with. In the past he has handled the problem by ignoring most of the available information and relating his judgments to scores on two or three measures in which he happens to have confidence.

It seems reasonable to believe that, if all of a student's valid past performance measures could be appropriately weighted and combined into a single statement of the probability of his success or failure, administrative decisions might become more accurate. Administrator knowledge of such probabilities should lead to the earlier dropping of men with high potentiality for later failure and to the retention of men who have good chances to complete the program. This, in turn, should result in marked improvement in the efficiency of utilization of training facilities.

Initial Procedure

The selection and training records of all men who entered naval aviation training during calendar 1959 were used as the basic data. The intercorrelations of the scores and grades shown in Table 1 were obtained in a succession of matrices which began with the selection measures. A new matrix was computed at each point that scores on one or more additional variables became available. In conjunction with each matrix, bi-serial correlations with a completed-failed dichotomy were computed for each measure. Following this, multiple correlations with the dichotomous criterion (Wherry-Doolittle) and beta weights were computed. Table 2 shows

TABLE 2

Variables and Multiple Validities for Predicting Failure at Successive Stages of Naval Aviation Training

Matrix	Variables	Variables in Multiple	N		R
			Completed	Failed	
1	1-6	1, 2, 3, 4, 6	670	243	.37
2	1-7	1, 2, 3, 4, 6	549	185	.38
3	1-12	1, 2, 3, 4, 6, 8, 9, 10	549	185	.42
4	1-13	1, 2, 3, 4, 6, 7, 8, 9, 10, 13	549	184	.46
5	1-17	1, 2, 3, 4, 6, 7, 8, 11, 13, 14, 16	513	155	.50
6	1-19	2, 3, 4, 6, 7, 8, 11, 13, 14, 16	513	153	.54
7	1-24	no change	513	153	.54
8	1-25	3, 4, 6, 8, 14, 15, 16, 18, 25	513	153	.56
9	1-29	no change	513	153	.56
10	1-30	3, 6, 9, 10, 11, 13, 14, 15, 16, 18, 25, 29, 30	502	116	.66
11	1-31	2, 3, 6, 8, 13, 14, 15, 16, 18, 25, 30, 31	502	83	.57
12	1-32	3, 6, 8, 13, 14, 16, 18, 25, 30, 32	464	64	.59
13	1-33	3, 6, 8, 14, 16, 18, 25, 30, 32, 33	464	55	.55
14	1-34	3, 8, 14, 16, 18, 25, 30, 31, 32, 34	441	44	.58

TABLE 3
Simplified Multiple Correlations with Completed-Failed Criteria

Stage	Variables in Multiple	R
1st and 2nd weeks	1, 2, 3, 4, 6	.38
3rd, 4th and 5th weeks	1, 2, 3, 4, 6, 8	.42
6th, 7th and 8th weeks	1, 2, 3, 4, 6, 8, 13	.45
9th and 10th weeks	1, 2, 3, 4, 6, 8, 13, 14, 16	.49
11th, 12th, 13th weeks	2, 3, 4, 6, 8, 14, 16, 18	.54
14th, 15th, 16th weeks and Pre-solo prior to hop #9	3, 4, 6, 8, 14, 15, 16, 18, 25	.56
Pre-solo, hop #9 and after, and T-34 Precision	3, 6, 13, 14, 16, 18, 25, 30	.65
Transition	6, 8, 14, 16, 18, 30, 31	.55
T/28 Precision	3, 14, 16, 18, 30, 32	.56
Acrobatics	3, 6, 8, 14, 16, 18, 30, 32, 33	.54
After Acrobatics	3, 8, 14, 16, 18, 30, 32, 33, 34	.57

the number of variables in each matrix and in each multiple correlation, and the multiple correlations with the completed-failed criteria. (Failure cases included flight failures, academic failures, and men dropped for disciplinary reasons.)

For the operational use of the prediction formulae certain simplifications were made. First, wherever two or three successive formulae were identical in variables and similar in weights, only one was used. Second, whenever the last several variables in a multiple added only three or four thousandths to the magnitude of the validity coefficients, these were dropped and the formulae re-computed. The variables in these simplified formulae, the periods of training to which they apply, and their respective multiple correlation coefficients with the criteria are shown in Table 3.

Initial Application

In order to transform these multiple validities into probability estimates that could be used to increase the accuracy of decisions about individual students the following steps were taken.

1. For each student who had entered training during 1959 a regression score was computed for each of the 11 stages of training shown in Table 3. Thus, at the first stage (1st and 2nd weeks), 1078 regression scores were computed by multiplying each student's scores by the appropriate weights and summing the products. This was repeated for each later stage for those students still in the program.
2. At each stage the frequency distribution of the regression scores

of men who subsequently graduated from training was compared with the frequency distribution of the scores of those men who subsequently failed to complete training. By dividing these distributions into five or six segments, and by determining the relative number of completing and noncompleting students within each segment, it was possible to establish empirical probabilities of success for each regression score.

3. The results were put into tables of variables, beta weights, regression scores, and probabilities of success for use at each level of training (see Fig. 1).

Formal instructions for obtaining a statement of the probability of success of individual students at any point in training were issued to the administrative personnel. These statements gave the odds for or against the student's subsequent completion of training and were obtainable by calling in the student's name, pre-flight class number, and certain recent grades to a "Student Prediction Center." Statistical clerks in the center computed the regression score for the particular stage of training and compared it with the appropriate Regression-Probability Table (as in Fig. 1). It was

Variables	Mean	S. D.	Weight
Spatial Apperception	20.60	5.48	.27
Biographical Inventory	33.20	8.20	.10
Education	6.74	1.22	.41
Peer Rating	50.30	10.86	.20
Navigation	48.92	6.80	.34
Trampoline	30.15	5.98	.17
Jump Reach (I)	11.03	2.77	.20
Engineering	49.97	8.25	.24
Physics	48.13	9.63	.10

Regression—Probability Table

Score	Percent	Will Complete	Won't Complete
60.0 or more	9	7	1
53.0-59.9	43	3	1
49.0-52.9	28	1½	1
45.0-48.9	14	1	1½
44.9 or less	6	1	4½
$N = 967$ $R = .56$			
No completions below 40.9—ten failures			

Figure 1. Table for determining probability of graduation—14th, 15th and 16th weeks.

usually possible to complete this operation in 10-15 minutes per student.

Administrators were cautioned that these probability statements were based upon all of the significant scores, grades, and ratings in the student's record, weighted in accordance with their relative importance to success, and that they were not to give additional consideration to any of the individual measures in the record. On the other hand, they were told that each case was to be decided on its unique merits, recognizing that many important aspects of a case of a man might not be reflected in this probability statement.

Revised Procedure and Crossvalidation

In the initial procedure the criterion was that of completed failure (flight, academic, or disciplinary), and the beta weights used were those that maximized the correlations with completely failed dichotomies. It was originally reasoned that cases that did not complete for causes other than failure (voluntary withdrawal, medical, etc.) would be randomly distributed on the continuum of completed-failed regression scores and that quite valid predictions of the empirical probabilities would come from using these scores.

Further study of the problem, however, indicated that slightly more useful (and more stable) predictions could be obtained by using as the criterion a completed-incompleted dichotomy, despite the fact that putting all types of attritions together lowered substantially the magnitudes of the obtained multiple correlation coefficients.

TABLE 4

Crossvalidation of Multiple Correlations with Complete-Incomplete Criteria

Week	No. Variables	No. Predictors	<i>R</i> 1959	<i>R</i> 1960
1	5	4	.305	.297
2	6	4	.305	.297
3, 4, 5	7	4	.317	.300
6, 7, 8	8	5	.336	.406
9, 10	11	6	.360	.410
11, 12, 13	12	6	.384	.448
14, 15, 16 and Pre-solo prior to hop #9	15	6	.407	.414
Pre-solo, hop #9 and after, and T-34 Precision	16	7	.465	.394
Transition	17	9	.424	.435
T/28 Precision and beyond	18	7	.450	.444

PRE-SOLO, AFTER HOP #9, and PRECISION (VT-1)

Slide 5137 and ask for "Student Prediction." Give the student's name, pre-flight
 er and his present point in training. Give his pre-solo flight grade to date.

name and phone number.

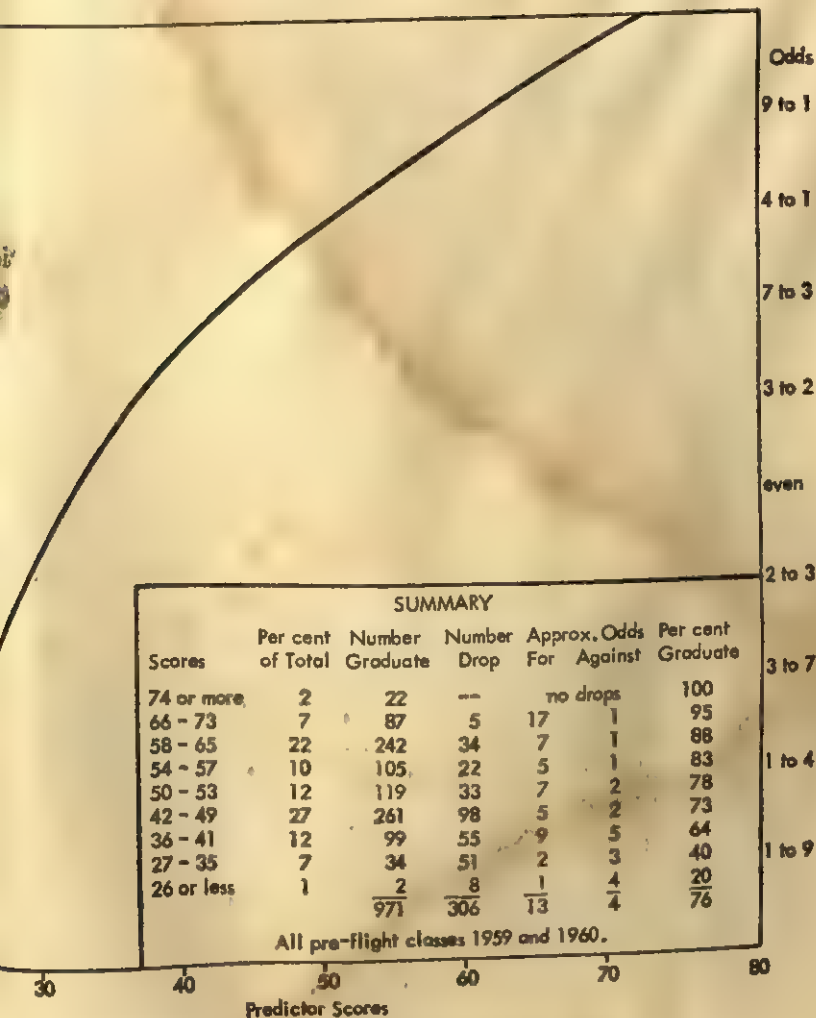


Figure 2. Graph-Table for interpreting predictor scores.

ts. Therefore, after eliminating those variables which had not
 a found useful in the original prediction formulae, the matrices
 e re-run on the 1959 students including all of the attrition cases.

The Wherry-Doolittle method was used to re-determine the multiple correlations and the beta weights appropriate to prediction at each stage. These formulae were used to compute regression scores at each stage on 976 students who entered in 1960. These scores were then correlated with the completed-incomplete records of the 1960 entrants. Table 4 shows the numbers of variables in each of the formulae, the R for the 1959 criterion at each stage, and the comparable R 's for the 1960 crossvalidation students. Confidence in the prediction procedures was reinforced by the lack of shrinkage in the crossvalidation coefficients.

Revised Application

Six months' experience with the student prediction system (as described earlier) had indicated that certain aspects of the system could be made more meaningful to the administrators using it. First, the regression formulae were modified so that distributions of regression scores had means of 50 and standard deviations of approximately 10—thus conforming to the grading system currently in use. Second, graph-tables such as the one shown in Figure 2 were prepared for each stage of training and distributed to those offices in which decisions about students must be made. This was done for two reasons: first, because some administrators appeared to have difficulty using the "odds" statements; and second, because the division of the distribution of regression scores into five or six relatively gross segments meant that there might be real differences in probabilities near the borders of the segments. The instructions were then modified so that when the administrator calls for a prediction on a student, he is given a "Predictor Score" which he can then interpret from the appropriate graph-table. These permit the administrator to use whichever method of interpreting he finds most comfortable.

ELECTRONIC COMPUTER PROGRAMS AND
ACCOUNTING MACHINE PROCEDURES

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District

and the

University of Southern California

<i>A FORTRAN Program for a Central Prediction System.</i> W. L. BASHAW	201
<i>Test Scoring and Item Analysis Programs.</i> ROBERT C. NICHOLS AND WILLIAM TETZLAFF	205
<i>An Application of Computer Programing to Test Analysis and Item Analysis.</i> WALTER DICK AND RICHARD E. SPENCE	211
<i>An IBM 1401 Computer Program for Item and Test Analysis.</i> ROBERT A. JONES, CALVIN PULLIAS, AND WILLIAM B. MICHAEL	217
<i>Test Scoring and Analysis with the Film-Optical Sensing Device for Input to Computers.</i> DOROTHY S. EDWARDS	221
<i>Polyserial Correlation Programs in FORTRAN.</i> NATHAN JASPEN	229
<i>A 1620 FORTRAN Program for Compiling a Flanders-Amidon Interaction Analysis Matrix.</i> ROBERT M. RIPPEY	235
<i>Computer Search for Group Differences.</i> M. CLEMENS JOHN-SON	239

IN view of the tremendous advances that have been made in the adaptation of electronic computers and accounting machines to the processing of statistical data, sections of the Spring and Autumn issues of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT are devoted to the publication of such programs as are appropriate to psychometric procedures. Programs relevant to such problem areas as factor analysis, item analysis, multiple regression procedures, the estimation of the reliability and validity of test pattern and profile analysis, the analysis of variance and covariance, discriminant analysis, and test scoring will be considered. Customarily a program should be expected not to exceed six or eight printed pages. Manuscripts of four or fewer printed pages are preferred. Each manuscript will be carefully reviewed as to its suitability and accuracy of content. In some instances an accepted paper may be returned to the author for possible revisions or shortening. The cost to the author will be fifteen dollars per page for regular running text. The extra cost of the composition of tables and formulas will be added to the basic rate. Manuscripts received up to November first will be considered for the Spring issue; manuscripts received between then and May first will be considered for the Autumn issue.

All correspondence should be directed to

William B. Michael

Professor of Education and Psychology
University of California, Santa Barbara
Santa Barbara, California

A FORTRAN PROGRAM FOR A CENTRAL PREDICTION SYSTEM¹

W. L. BASHAW

Florida State University

THE program described is based primarily on Tucker's prediction model for central prediction systems (Tucker, 1963). The actual model is an adaptation of Tucker's basic model (Bashaw, 1963).

The term "central prediction system" means a statistical scheme for predicting the success in any one of several colleges of graduates of any one of several schools in a given school system on the basis of school achievement, test scores, and school attended. The central prediction technique differs from customary regression methods in several respects. A single analysis yields predictions for all alternative courses of action open to students in the system. Data from all schools and colleges in a school system can be included in the analysis. Differences among institutional grading practices can be taken into account and corrections for these differences can be calculated.

An assumption basic to Tucker's central prediction technique is that the distributions of predictor achievement measures and criterion achievement measures are not the same among institutions. Similarly, the relationships between predictor achievement measures and criterion achievement measures are not the same among institutions. The model corrects for this heterogeneity by the use of institutional parameters.

The Prediction Model

The independent variables of the model are constructed from the knowledge of school achievement, school attended, and test scores.

¹ The development of this program was partially supported by a National Science Foundation grant to the Florida State University Computing Center (Grant No. GP-671).

Consider a school transfer system in which there are ns schools and nc colleges. Data for each subject include scores on nt tests, a school grade average, and a college grade average (criterion).

The independent variable vector, X , has nx elements, where $nx = 2*ns + nt + 2$. The first $2*ns$ elements consist of pairs of values that correspond to the ns schools. If a particular subject attended, say school four, then the seventh element has a value equal to the subject's school grade average and the eighth element has the value one. All other values of the first $2*ns$ elements are zero. The next nt elements consist of scores on the nt tests. The $2*ns + nt + 1$ element consists of the school grade for each subject, while the last element is unity for each subject.

An examination of the independent variable vector will reveal terms corresponding to a main effect of schools, a main effect of tests, a main effect of school grades, and an interaction effect of schools by school grades.

The elements of the predictor vector, X , are related to the criterion through two matrices of weights, denoted by A and W . The matrix A , when applied to the data vector, X , defines np independent predictor composites. Thus, the matrix A serves as a factor matrix as it is used to reduce X to np independent variables. The value of np is less than nx and is at a maximum equal to nc . It is assumed that the elements of A are the same for all colleges. The estimates of the elements of A are based on data from all schools and colleges. For each college c , there being one vector for each college. The W_c are the usual least squares estimates of regression weights and are based on all data from college c . The final equation for predicting a grade at college c is $\hat{Y} = XAW_c$.

The solution to the estimation problem is given by Tucker (1963). His solution is based on an initial estimate of the A matrix. (This estimate can be quite crude and still be satisfactory. A suggested initial estimate of A is the use of regression weights obtained from a standard multiple regression solution, where the weight for school grades is used for each school.) The initial estimate of A is applied to the X vector to form initial estimates of the predictor composites. The regression of the criterion on these predictor composites is used to formulate the usual least squares regression estimates for each college (W_c). The obtained W matrix is used to calc-

culate a new estimate of A . This cycle is continued until there is no change in the prediction system.

Program Input

1. Data format
2. Columns 1 and 2 — ns (number of schools)
Columns 3 and 4 — nc (number of colleges)
Columns 5 and 6 — nt (number of tests)
Columns 7 and 8 — np (number of columns in the original estimate of A)
3. Columns 1 and 2 — 01 (index for school 1)
Columns 3 and 4 — 01 (index for college 1)
Columns 5, 6, 7, 8 — number in cell
4. Data card—according to variable format, but in the order:
test scores, school grade, college grade.
There is one data card for each subject. Following the last data card of a cell, a type 3 card is read for the next cell. The indices should be stepped up for each cell. Continue for each school-college cell.
5. Columns 9, 10, 11, 12 — 9999 (indicates the end of data).
6. Read the estimate of A by rows in the format $XX.XX$. There should be nx cards, one for each row.

Program Output

1. College and total sums of variables and standard deviations.
2. Sample sizes.
3. The original estimate of A and the original estimate of W .
4. The number of roots for each iteration.
5. The multiple correlations for each college for each iteration.
6. The final W vector for each college.
7. The final A for standardized variables.
8. The final A for raw scores.
9. The regression equation constant for each college.

Program Limitations

The program requires a 32K computer with the ability to handle chain jobs. The present program is a three link chain job and uses tape drive A4.

The program is limited to six colleges, twelve schools, and five test scores. Considerable storage space is still available when the chain job program is used. Therefore larger problems can be handled by changing dimension statements or by increasing the number of links.

The program stops iterating when the improvement in each multiple correlation is no more than 0.0001. In addition, a programmed halt will occur if more than 25 iterations are required.

REFERENCES

- Bashaw, W. L. "A Central Prediction System for Predicting the Success of Junior College Transfers in Florida Universities." Unpublished Ph.D. dissertation, Florida State University, Tallahassee, Florida, 1963.
- Tucker, L. R. *Formal Models for a Central Prediction System*. Richmond, Virginia: William Byrd Press, 1963.

TEST SCORING AND ITEM ANALYSIS PROGRAMS¹

ROBERT C. NICHOLS

National Merit Scholarship Corporation
and

WILLIAM TETZLAFF

Northwestern University

ITEM analysis for the development of new tests and scales is difficult, not because the calculations are complicated, but because the clerical task is immense. A typical scale construction project involves (a) tallying item responses of several hundred subjects to several hundred items in high and low criterion groups, (b) obtaining phi coefficients for each item, (c) scoring a provisional scale, (d) tallying the items again for high and low groups on the provisional scale, (e) obtaining phi coefficients, and (f) scoring the revised scale on a cross-validation group. For some purposes additional tallying and scoring is required.

To reduce the clerical labor of scale construction, computer programs were written to perform the tallying and scoring operations. These programs, which were written in *FORTRAN IV* and *MAP*, are currently set up to run on the IBM 709 or 7090 *IBSYS* Monitor System.

Data Preparation

Item response data are punched into cards in any format that is convenient. More than one item may be punched in a single card

¹ These programs were written as part of the research program of the National Merit Scholarship Corporation which is supported by grants from the National Science Foundation, the Carnegie Corporation of New York, and the Ford Foundation.

column, and the programs are most efficient when as many items as possible are punched on a single card.

Item responses can be economically punched from an answer sheet by punching the first two answer sheet columns as 0 and 1 (for dichotomous items), the next two as 2 and 3 on a second card, the next as 4 and 5 on a third card, and so forth. After all cards have been punched and edited, they are gang punched for each subject into a single card. If an "x" punch is placed at the end of each answer sheet column, cards on which an item has been omitted or punched twice can be sorted out, and a satisfactory degree of accuracy can be obtained without key verification. Through use of this method, *California Psychological Inventory* answer sheets (480 dichotomous items) were punched commercially for 35 cents each. Alternative procedures, which may be more economical for large samples, include mark sense cards and the document reader.

For both tallying and scoring, the data are loaded on a separate data tape, each card as a 28 word binary record. A special IBM 1401 program is used.

Scoring

Data. Up to five data cards per subject are allowed. Data for several problems may be loaded successively on the tape, and in any given run the program can be instructed to skip those blocks of data which are not to be scored. Thus, all the data for a given project can be stored on tape, and any block of data can be scored for new keys at any time.

The first card in each block of data should contain 80 columns of alphameric identification. This card will not be scored, but will be printed and punched to identify the output. Dummy subjects may also be included with appropriate alphameric identification in the *ID* field in order to yield information about the keys. For example, a dummy subject with all "true" responses will give as scores a count of the number of items keyed "true" in each key.

The program assumes a four-digit *ID* number in columns 3-6 of the first card for each subject. The *ID* field can be changed easily in the FORTRAN program to any number of columns in any desired card field.

Keys. Up to 30 keys for each block of data can be scored in a single run. A key consists of from one to five cards (one for each

data card per subject) with punches corresponding to the item responses to be counted. The score for a given subject is the number of punches in his data cards which have corresponding punches in the key cards. The keys are loaded on a separate key tape in binary with the same 1401 program used to load the data. Keys for additional problems are loaded successively on the tape.

Control Cards. The program is loaded on the *IBSYS* input tape followed by a problem card and cards specifying the punch and print output format.

Problem Card: Cols. 1-2 Number of keys.

- 3 Number of cards per key (or the number of cards per subject).
- 4 Number of punch format cards.
- 5 Number of print format cards.
- 6-7 Number of blocks of data to be skipped before scoring.
- 8-79 Alphameric title to be printed on output.

Output. The scores are written on tape for off line printing and punching according to the format specified in the print and punch format cards. The format should specify integer fields except for the *ID* number which should be composed of single character *A* fields. For example (2X, 4A1, I3, 12I2). The *ID* number is the first number printed and punched followed by the scores in the order in which the keys appeared on the key tape.

Operation. The monitor system provides instructions for hanging the various tapes. The program skips the specified number of blocks of data, prints and punches the first card in the block as identification, reads in the data for the first subject, scores all keys, writes the scores on both print and punch tapes, and reads in the data for the next subject. This method of operation continues until the end of file mark following each block of data is reached. The program then skips the number of blocks of data specified on the next problem card, prints and punches the initial identification card, and scores the next block of data.

Time. In one minute on the 709 between 1000 and 2000 subjects, depending on the number of keyed responses, can be scored for one single card key.

Tallying

There are two forms of the program for tallying. Form 1 reads in two groups of item response data cards as presented by the user, tallies all requested card positions, and prints for each card position N 's and percentages for each group as well as ϕ and ϕ/ϕ_{\max} coefficients, and the item content. The same data are then printed again with the items arranged in order of the size of ϕ . Form 2 of the tallying program computes and prints the same statistics, but instead of taking groups sorted by the user, the program forms its own groups on the basis of continuous score data for the subjects. Both forms tally only one item response card per subject. Multiple cards per subject should be run as separate problems.

Form 1

Data. Item response data cards are sorted into two groups and loaded in binary on a separate data tape. Each group should be followed by a card with "END" punched in columns 1-3. Data for additional problems may follow successively on the tape.

Control Cards. The following cards should follow the program in the order described:

Title card: Cols. 1-80	Alphameric identification to be printed at the top of each page of output.
Group name cards: Cols. 1-12	Alphameric group names (one card for each group).
Item name cards: Cols. 1-3	Card position number (card positions are numbered consecutively beginning with the "y" punch in column 1, down column 1, then down column 2, then down column 3, and so forth. The "9" punch in column 80 is card position 960).
Cols. 5-80	Alphameric item name

The deck of item name cards should be followed by a blank card. Only those card positions for which there are item name cards will be tallied. For subsequent problems a single card with 999 in columns 1-3 may be substituted for the deck of item names and the

blank card, in which case the item names from the previous problem will be used. The title card for the next problem should follow the item name deck of the preceding problem.

Form 2

Data. Two types of data are necessary for this form of the tallying program: item responses to be tallied and continuous variables which are the basis for forming groups.

The item response cards should be loaded on a separate data tape in binary in ID order. The ID field is assumed to be columns 3-6, but this can be easily changed in the FORTRAN program. The last card should have "END" as the first three columns of the ID field (not necessarily the first three card columns as in Form 1).

The continuous variable cards contain the scores on which groups are to be formed by the selection of high and low scoring subjects. High and low groups are tallied separately for each continuous variable. A maximum of 50 variables is allowed. The first continuous variable should be an ID number corresponding to the ID numbers of the item response cards. The last continuous variable card should have "END" as the first three characters of the ID field. The continuous variable cards should be in ID order and follow the program and control cards on the monitor input tape.

Control Cards. The following control cards follow the program in the order described and precede the continuous variable data.

Title card: Cols. 1-80

Alphanumeric identification to be printed on the first line of each page of output.

Problem card: Cols. 1-3

Number of continuous variables (not including the ID number).

4-6 Estimate of the number of item response cards (this is used for efficient tape handling and does not have to be exactly correct).

7 Number of variable format cards specifying format of continuous variable cards.

Variable format card(s):

Standard FORTRAN variable format cards. The first four fields

should be 4 fields for the four digits of the ID number. The remaining fields should be integer fields. For example: (4A1, 1013).

Subproblems. Following the continuous variable data there should be a group of control cards for each subproblem. A subproblem is a tallying of high and low groups formed on the basis of one of the continuous variables. There may be more than one subproblem for each continuous variable. Each subproblem should have the following control cards in the order described.

Subproblem card: Cols. 1-2 A number less than 51 indicating the continuous variable to be used for group formation.

3-7 The lowest score permitted in group 1 (the high group).

8-12 The highest score permitted in group 2 (the low group). (Note that according to the contents of col. 3-12 a subject may be tallied in *either*, *neither*, or *both* groups.)

Secondary Heading cards:
Cols. 1-80

Alphanumeric identification to be printed as the second line of each page of output for this subproblem.

Group name cards: Cols. 1-12 Alphanumeric group name (one for each group).

Item name cards: Item name cards or 999 card (same as for Form 1).

Operation. The program reads the continuous variable cards and the item response cards, compares the ID numbers, and writes the data for matching cards on tape. This tape is read for each subproblem and the item responses tallied in the appropriate group.

Time. Both forms of the tallying program will tally all 960 card positions for two groups of 250 item response cards each in about five minutes on the 709.

AN APPLICATION OF COMPUTER PROGRAMING TO TEST ANALYSIS AND ITEM ANALYSIS

WALTER DICK AND RICHARD E. SPENCER¹

The Pennsylvania State University

SINCE the Fall of 1963, the Office of Examination Services of the University Division of Instructional Services at the Pennsylvania State University, has been using the facilities of the University's Computation Center to process test results and analyses. The major purpose of such computations is to attempt to improve instructional effectiveness through the improvement of examinations and examination systems. During the initial year of computer-assisted operations, over 60,000 test papers were scored and over 20,000 were submitted for item analyses.

A system of programs has been developed which uses either key-punched data cards or cards punched by the Digitek 100A Optical Scanner Test Scoring Machine. Tests with up to 150 items are processed daily on the University Computation Center's IBM 1401-7074 system. Results are usually available to instructors within 12 to 24 hours.

Services Available

When an instructor requests that his tests be scored by Examination Services, he is provided with a supply of answer sheets which have been specially designed for the Digitek Machine. The Digitek scores the sheets at a rate of 2520 per hour. When the punch unit is in operation (punching out identification, test and class code, item responses, total score and card sequence), the rate is 1260 sheets per hour. These punched cards can then be used in the Total Test Pro-

¹ Now with the Office of Instructional Resources, University of Illinois.

gram, Summary Program and/or Item Analysis Program described below.

Computer Programs

Total Test Program

Data cards containing optional identification, such as student number or name, and test score are used in this program. There is no limit to the number of students' scores which may be used. The output includes the following information.

1. Identification (course, instructor, date, number of students, number of items).
2. Double column listing of student names or numbers, test raw scores, and associated standard score ($\bar{X} = 500$; $\sigma = 100$).
3. List of absentees (obtained by matching class cards with answer sheets).
4. Table of raw score—standard score—percentile equivalents.
5. Graphic frequency distribution of raw scores.
6. Statistical summary of the test data (mean, variance, standard deviation, range, standard error of measurement, standard error of the mean, Kuder-Richardson 21 reliability estimate, skewness, and kurtosis).

When the first test of a term is processed for an instructor, the total test program produces summary cards for all students which include identification and either their raw score or standard score. These cards are resubmitted with each subsequent test processed during the term (in addition to the test data cards). Thereby an up-to-date summary of all students' scores on all tests is provided.

Summary Program

A Pennsylvania State University ruling requires that all student grades must be filed with the Registrar's Office within 48 hours after the final examination. In order to provide instructors with sufficient time to assign letter grades to the students, a special version of the total test program has been devised. The summary program or SUM 2 is used to process the data from the last course examination and the up-dated summary cards.

The program provides all the information listed above under

"total test." The computer then scans the summary scores and prints out the names and scores of those students who do not have the minimum number of tests (as stipulated by the instructor) to qualify for the summary procedure.

A number of options are available as to how the scores will be summarized. These include:

1. a simple mean of all the scores for the term for each student;
2. a rank ordering of scores with a mean of the x highest scores for each student;
3. "holding in" of up to three scores (e.g., mid-term, final), rank ordering the remainder, taking the x highest of these scores and adding in the "held in" scores and obtaining a mean; and
4. multiplication of each score by a stated constant (e.g., .30 times the first score, .30 times the second, and .40 times the third).

The printout of the Summary Program includes the regular total test printout plus another heading and identification of the summary procedure used. For each student the following is printed.

1. Identification
2. Score which results from summary procedure
3. Converted score equivalent of score obtained in 2
4. Percentile rank
5. Test scores upon which the summary was based.

The program also produces a statistical summary of the scores produced by the summary procedure. This summary includes only the mean, variance, standard deviation, and range. Punched cards are also produced for each student which include all his test scores and his summary score.

Item Analysis

The data used in the item analysis program must include a total score and item responses for each student. Digitek-punched cards used in the Total Test Program may also be used for item analyses; however, the program, at the present time, is limited to 150 items per run. If key-punched cards are used, the number of items which can be processed is approximately 1000. There is no limit on the number of students included in the run.

TABLE 1
Format for Item Analysis Printout

Item Number 1						Correct answer is 3
Responses	Lowest Fifth	Second Fifth	Middle Fifth	Fourth Fifth	Highest Fifth	Response Total
Omit	4	4	1	1	0	10
1	5	3	4	5	1	18
2	9	5	3	5	0	22
3	7	13	14	19	25	78
4	6	7	9	2	5	29
5	0	0	0	0	0	0
Total	31	32	31	32	31	157

Proportion of total group of 157 students answering correctly = 0.497.

Correlation between success on this question and total score on test = 0.540.

Point-biserial correlation = 0.431, $T = 5.944$.

Mean score SS correct = 35.12.

Mean score SS incorrect = 25.77.

Table 1 is a sample of a printout of an item evaluation. It includes:

1. Item number
2. Correct answer
3. Number of students selecting each response and number omitting the item
4. Responses selected by up to 5 subgroups within the total sample. (In the example in Table 1, the students were divided into fifths on the basis of their total scores.)
5. Difficulty index (proportion of students getting the item correct)
6. Biserial correlation coefficient
7. Point-biserial correlation coefficient
8. t associated with the point-biserial coefficient
9. Mean total score of the student who had the item correct
10. Mean total score of the students who had the item incorrect.

An optional feature, not shown in Table 1, is a letter grade evaluation of the item difficulty, the item discrimination, and of the item as a whole.

An analysis such as appears in Table 1 is printed-out for each item on the test. At the end of the analysis of all the items a summary is printed which includes the following.

1. Mean difficulty of the test items

2. Mean biserial correlation
3. Standard error of the biserial coefficients
4. Estimated interitem correlation
5. Kuder-Richardson 20 reliability estimate
6. Test mean, variance, and standard deviation
7. Frequency distribution of the item difficulties
8. Frequency distribution of the item biserials

Value of the Programs

The total test and item analysis programs have been placed in the permanent tape library of the Pennsylvania State Computation Center. The staff of Examination Services are required to submit only the data cards for a computer run, along with a call card for the use of a particular program. An indication of the speed with which these programs operate can be gained from the following sample times.

Three total tests, each with 100 items, with a total of 652 students required one and one half minutes.

An item analysis with 50 items and 150 students required one minute.

An item analysis with 50 items and 350 students required one and a half minutes.

The value of these programs is reflected in a number of ways. Objective testing procedures have been emphasized throughout the University, instructors and graduate students have been spared many hours of clerical work, and conferences with instructors about their test results lead to an increased awareness of measuring and grading procedures and the need for better evaluative procedures.

(Sample printouts and FORTRAN listings are available from the first-named author.)



AN IBM 1401 COMPUTER PROGRAM FOR ITEM AND TEST ANALYSIS

ROBERT A. JONES AND CALVIN PULLIAS

University of Southern California

AND

WILLIAM B. MICHAEL

University of California, Santa Barbara

It is the purpose of the writers to describe an IBM 1401 Computer program for item and test analysis that has been found to be particularly useful for multiple-choice examinations of 75 or fewer items. The program has afforded a means for college instructors to obtain very rapidly information about item and test characteristics that has been highly useful both in the evaluation of student performance and in the editorial revision and modification of individual test items.

Characteristics of the Program

Scope. In addition to furnishing a score on the test for each examinee, the program prints the following information: (1) the number of questions correct, the number wrong, and the number omitted, the formula score, and a transformed score (i.e., a Z score with a mean of 50 and a standard deviation of 10)—all these data being opposite the examinee's identification number; (2) a frequency distribution of the scores for the sample; (3) item response information (i.e., frequency of responses to each of the alternatives as well as frequency of omits) for upper-half and lower-half criterion groups or for the subgroups answering correctly or incorrectly; (4) either a phi coefficient or a biserial and point biserial coefficient; (5) an item difficulty value in terms of the proportion of correct re-

sponses (either corrected or uncorrected for chance success); (6) both Kuder-Richardson Formula 20 and Formula 21 estimates of reliability; (7) the mean and standard deviation of the total score distribution. The program also allows the use of a two digit external criterion score in place of an internal criterion. Sense switch options exist that permit the correction or the lack of correction of scores and item difficulties for chance success.

Limitations. The routine was designed to use the data punching capability of the IBM 1230 Optical Scanner. In order that a maximum "through-put" could be obtained, the IBM 1230 was programmed to punch in packed form. A separate IBM 1401 routine (which is also available) was used to unpack the IBM 1230 output. The original version of the test analysis routine will accept 75 or fewer items. (A modified deck is available in which the number of items may be extended to 150, but the use of an additional tape drive is required.) The number of subjects must be equal to or fewer than 999.

For relatively short tests the computing time T in seconds is approximately $T = .60Q + .20N + 5$, where Q is the number of questions and N is the number of examinees. For example, 75 items and 60 subjects require 62 seconds. The total time needed by the IBM 1230 to punch the data on 75 questions for 60 subjects is approximately 4 minutes. About 30 seconds are necessary for the IBM 1401 to unpack the 1230 output.

Input. Input is by cards only. The detailed deck set-up is described in a subsequent section concerned with the preparation of the job.

Output. Output is printed in essentially tabular form and is appropriately labelled.

Machine requirement. An IBM 1401 with 8K storage and special features is required. A tape unit is used—a unit which must be called Tape "2."

Preparation of the Job: Deck Set-Up

Between the object deck and the data deck the following six control cards must be inserted.

Label cards (first, second, third, and fourth) Although these cards may be left blank, whatever is punched in them will be printed as identifying information on the first four lines of the output.

Data control card

Column 1 corresponds to the number of choices in items (the maximum being five).

(fifth)

Column 2 and column 3 are employed for the number of questions—column 2 for the tens digit and column 3 for the units digit. The maximum number of questions is 75.

Columns 4 and 5 are left blank.

Columns 6 and 8 may be employed on an optional basis for the number of subjects.

Column 9 is used as the basis for printing of data. The "one punch" is employed for printing of raw data. If no data printing is desired, the column is left blank.

Column 10 is used for criterion data. The "one punch" is employed for an external criterion. If the total score constitutes the (internal) criterion, the column is left blank.

Key card
(sixth)

Columns 1-75—each column contains the correct answer for the question corresponding to the column number (unused columns being left blank).

Columns 76-80 must be punched with five zeros.

Data cards

Columns 1-75 contain the student's response to each question (an omitted question being recorded as a blank or a zero).

Columns 76-77 are for a two digit external criterion if it is used.

Columns 78-80 are for student or card identification.

Availability

A listing of the program will be supplied free of charge. However, source decks will be supplied at a nominal charge to cover postage and handling.



TEST SCORING AND ANALYSIS WITH THE FILM-OPTICAL SENSING DEVICE FOR INPUT TO COMPUTERS

DOROTHY S. EDWARDS
American Institute for Research

Background

IN 1961 the American Institute for Research was engaged in a project for Headquarters, Air Force Reserve Officer Training Corps, which required the scoring and item analysis of two answer sheets for each of 15,000 cadets. The usually available techniques for accomplishing this, including the IBM 805, Graphic Item Counter, and key punching, were explored. These procedures were either too time consuming with the equipment and personnel available or too expensive. Therefore, the Census Bureau was approached about the possibility of using their Film-Optical Sensing Device for Input to Computers (FOSDIC) machine, and the Bureau agreed to try this application of the process.

The FOSDIC Operation

The FOSDIC operation has been described fully in a paper by McNelis, (1961). Basically, the procedure involves the recording of data on a FOSDIC document, or answer sheet, which is then microfilmed. These negatives are then scanned, frame by frame, by a controlled beam from a cathode ray tube. Where the marked responses have created a transparent spot in the negative, the beam activates a photo-electric cell and thus generates an electrical voltage. A plugboard is wired to designate a specific character to be placed on tape in a specific order to represent a marked answer. The

resulting magnetic tapes are then in a form usable with electronic computers.

The optical scanning device reads the lightest spot on the negative for any given set of options. This spot, of course, corresponds to the darkest mark made by the recorder, or examinee. This optical procedure overcomes the difficulty sometimes encountered with poor erasures, but introduces a further problem when multiple answers are acceptable. This difficulty too can be overcome with additional instructions to the machine. There are about 45 instructions which can be wired into the FOSDIC plugboard; some of them direct the movement of the electron beam geometrically from place to place on the microfilm frame, and others direct FOSDIC to read, to count, or to make conditional transfers.

FOSDIC's Performance

In order that the procedure might be adapted to the standard IBM answer sheet for this study, it was necessary for the Census Bureau personnel to build a special jig to hold the answer sheet and to add FOSDIC's sensing, or reference, marks. These sensing marks, which are small black squares or rectangles appearing on the document, are used by FOSDIC to align itself properly. Minor difficulties in the document, such as shrinkage, or tilt during microfilming, will be automatically corrected before the film is scanned, because FOSDIC seeks out the sensing marks, lines itself up on them, and, if necessary, adjusts the entire geometric pattern of its scan.

The sensing marks were applied to the Plexiglass cover of the jig which held the IBM answer sheet. The procedure required individual insertion of each answer sheet into the jig and the covering of it with the Plexiglass, an operation rather like inserting the answer sheet inside the cover of a book. The operators were not able to carry out this procedure as quickly as they could film the usual Census documents which were spiral bound and required only the flipping of a page, taking the picture, flipping the page, and so forth. The regular document used in the 1960 census permitted photographing two pages at a time—another short-cut which was not available. It is estimated, however, that with a regular FOSDIC-tailored answer sheet, similar to that reproduced in Figure 1, a single camera could photograph over 2400 frames per eight-hour period. If the "double photography" technique were employed, 4800

TEST ANSWER SHEET FOSDIC

Each answer sheet will be considered an negative statement as that your name and the preprinted blank spaces will appear on other files.

The answers on this sheet will then be scanned by an electronic beam which will show through the clear film to create an image on magnetic tape. This magnetic tape is then used on an electronic computer which can compare your answers with the correct answers.

For your benefit, please observe the following rules:

1. Do not use ink. The machine can read ink, but you cannot even ink clearly enough.

2. A number 25 black pencil is best, but a number 3 or number 3 can be used. Anything softer than number 2 will cause problems, and anything harder than number 3 will not give a black mark.

3. If you make a mistake, erase clearly otherwise the wrong mark might be read.

For your identification, write your Social Security Number and Test Number in the appropriate boxes, and also fill in the proper marking area.

If you do not have a Social Security Number, enter zeros in all of the boxes and fill in the zero item of the marking area. Do not enter 00000 for Test Number.

SOCIAL SECURITY NUMBER

Write number in boxes and fill marking area.

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
1	2	3	4	5	6	7	8	9	0
2	3	4	5	6	7	8	9	0	1
3	4	5	6	7	8	9	0	1	2
4	5	6	7	8	9	0	1	2	3
5	6	7	8	9	0	1	2	3	4
6	7	8	9	0	1	2	3	4	5
7	8	9	0	1	2	3	4	5	6
8	9	0	1	2	3	4	5	6	7
9	0	1	2	3	4	5	6	7	8

APPLICANT'S TEST NUMBER

Write number in boxes and fill marking area.

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
1	2	3	4	5	6	7	8	9	0
2	3	4	5	6	7	8	9	0	1
3	4	5	6	7	8	9	0	1	2
4	5	6	7	8	9	0	1	2	3
5	6	7	8	9	0	1	2	3	4
6	7	8	9	0	1	2	3	4	5
7	8	9	0	1	2	3	4	5	6
8	9	0	1	2	3	4	5	6	7
9	0	1	2	3	4	5	6	7	8

0000000000

1 2 3 4 5

1	A	B	C	D	E	31	A	B	C	D	E	61	A	B	C	D	E	91	A	B	C	D	E	121	A	B	C	D	E
2	A	B	C	D	E	32	A	B	C	D	E	62	A	B	C	D	E	92	A	B	C	D	E	122	A	B	C	D	E
3	A	B	C	D	E	33	A	B	C	D	E	63	A	B	C	D	E	93	A	B	C	D	E	123	A	B	C	D	E
4	A	B	C	D	E	34	A	B	C	D	E	64	A	B	C	D	E	94	A	B	C	D	E	124	A	B	C	D	E
5	A	B	C	D	E	35	A	B	C	D	E	65	A	B	C	D	E	95	A	B	C	D	E	125	A	B	C	D	E
6	A	B	C	D	E	36	A	B	C	D	E	66	A	B	C	D	E	96	A	B	C	D	E	126	A	B	C	D	E
7	A	B	C	D	E	37	A	B	C	D	E	67	A	B	C	D	E	97	A	B	C	D	E	127	A	B	C	D	E
8	A	B	C	D	E	38	A	B	C	D	E	68	A	B	C	D	E	98	A	B	C	D	E	128	A	B	C	D	E
9	A	B	C	D	E	39	A	B	C	D	E	69	A	B	C	D	E	99	A	B	C	D	E	129	A	B	C	D	E
10	A	B	C	D	E	40	A	B	C	D	E	70	A	B	C	D	E	100	A	B	C	D	E	130	A	B	C	D	E
11	A	B	C	D	E	41	A	B	C	D	E	71	A	B	C	D	E	101	A	B	C	D	E	131	A	B	C	D	E
12	A	B	C	D	E	42	A	B	C	D	E	72	A	B	C	D	E	102	A	B	C	D	E	132	A	B	C	D	E
13	A	B	C	D	E	43	A	B	C	D	E	73	A	B	C	D	E	103	A	B	C	D	E	133	A	B	C	D	E
14	A	B	C	D	E	44	A	B	C	D	E	74	A	B	C	D	E	104	A	B	C	D	E	134	A	B	C	D	E
15	A	B	C	D	E	45	A	B	C	D	E	75	A	B	C	D	E	105	A	B	C	D	E	135	A	B	C	D	E
16	A	B	C	D	E	46	A	B	C	D	E	76	A	B	C	D	E	106	A	B	C	D	E	136	A	B	C	D	E
17	A	B	C	D	E	47	A	B	C	D	E	77	A	B	C	D	E	107	A	B	C	D	E	137	A	B	C	D	E
18	A	B	C	D	E	48	A	B	C	D	E	78	A	B	C	D	E	108	A	B	C	D	E	138	A	B	C	D	E
19	A	B	C	D	E	49	A	B	C	D	E	79	A	B	C	D	E	109	A	B	C	D	E	139	A	B	C	D	E
20	A	B	C	D	E	50	A	B	C	D	E	80	A	B	C	D	E	110	A	B	C	D	E	140	A	B	C	D	E
21	A	B	C	D	E	51	A	B	C	D	E	81	A	B	C	D	E	111	A	B	C	D	E	141	A	B	C	D	E
22	A	B	C	D	E	52	A	B	C	D	E	82	A	B	C	D	E	112	A	B	C	D	E	142	A	B	C	D	E
23	A	B	C	D	E	53	A	B	C	D	E	83	A	B	C	D	E	113	A	B	C	D	E	143	A	B	C	D	E
24	A	B	C	D	E	54	A	B	C	D	E	84	A	B	C	D	E	114	A	B	C	D	E	144	A	B	C	D	E
25	A	B	C	D	E	55	A	B	C	D	E	85	A	B	C	D	E	115	A	B	C	D	E	145	A	B	C	D	E
26	A	B	C	D	E	56	A	B	C	D	E	86	A	B	C	D	E	116	A	B	C	D	E	146	A	B	C	D	E
27	A	B	C	D	E	57	A	B	C	D	E	87	A	B	C	D	E	117	A	B	C	D	E	147	A	B	C	D	E
28	A	B	C	D	E	58	A	B	C	D	E	88	A	B	C	D	E	118	A	B	C	D	E	148	A	B	C	D	E
29	A	B	C	D	E	59	A	B	C	D	E	89	A	B	C	D	E	119	A	B	C	D	E	149	A	B	C	D	E
30	A	B	C	D	E	60	A	B	C	D	E	90	A	B	C	D	E	120	A	B	C	D	E	150	A	B	C	D	E

Figure 1
General Purpose FOSDIC
Answer Sheet

forms per day could be filmed. Scanning of the film by FOSDIC can be done at about one hundred answer sheet sides per minute. Thus, one camera day equals about one FOSDIC hour. Even with the encumbrances necessary to photograph IBM answer sheets, the job was completed within one month, and at a cost equal to one-third of the lowest card punch bid received.

For operations where the document is tailored to the FOSDIC requirements, preparation of the tapes would undoubtedly be faster and cheaper. Experience with FOSDIC in the national population census at the Bureau of the Census has indicated that FOSDIC is about 100 times as fast as card punching, and its error rate is less than one-half of one percent.

FOSDIC's Future

FOSDIC's role in compiling Census data is well established. There are in existence four machines, all located at the Census Bureau. The Bureau, which is currently exploring additional ways of using FOSDIC, has had success in completing plans and a trial run for putting its own personnel records onto FOSDIC forms suitable for automatic data processing. Figure 2 shows part of the type of document used for this purpose.

The present writer feels that FOSDIC also has great potential in the preparation of test results for Automatic Data Processing. It has enormous accuracy; some of its accuracy levels can be pre-set in such a way that the machine will give indication when it has fallen below the maximum tolerable error-rate. It is fast; it is cheaper than other presently available processes. The Census Bureau is now developing a set of standard answer sheets, a sample of which appears in Figure 1. A general purpose document has also been developed, in which the questions can be reprinted to align with the pre-printed answer spaces. Many other variations are possible, since FOSDIC has tremendous latitude in the placement of answer spaces. It can read light pencil, ink, colored pencil—anything that will photograph. The Census Bureau is also working on a procedure that will permit FOSDIC to read letters or numerals. A sample of such an experiment is shown in Figure 3. It is based on an exaggeration of the natural tendency for each numeral to extend beyond a basic square at one or more points in a unique fashion. FOSDIC also has latitude in the size of the document it can read. The largest

INSTRUCTIONS TO CARD PUNCH OPERATORS

- Punch cards only when CHANGE is indicated.
→ Punch items 1 and 2 into code. 1-15 for all 3 cards.

Fill one circle on each line

CHANGE NO CHANGE
Punch Do not punch

NAME

ADDRESS

POSITION TITLE

1. ORGANIZATION						2. EMPLOYEE'S SOCIAL SECURITY NO.															3. NATURE OF ACTION										4. CSC OR OTHER LEGAL AUTHOR.										5. REMARKS									
Office or Division Branch, section, or unit																					Code No:										Code No:										Code No:									
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

12. DATE OF APPOINTMENT AFFIDAVITS				13. POSITION NUMBER				14. SCHEDULE OR TYPE OF PAY PLAN FOR POSITION				15. SERIES CODE				16. GRADE				17. STEP				18. COMP. LEVEL				19. ADMINIST. TITLE				20. POSITION INFORMATION				21. APPOINTMENT			
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

30. TYPE OF APPOINTMENT				31. EXTENT OF DUTY				32. TYPE OF LIMITATION ON APPOINTMENT				33. NTE DATE LIMIT ON APPOINTMENT				34. SALARY/EARNINGS LIMIT ON APPOINTMENT				35. DAYS LIMIT ON APPOINTMENT				36. HOURS LIMIT ON APPOINTMENT				37. DATE			
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

38. IS EMPLOYEE ON DETAIL				39. EFFECTIVE DATE OF DETAIL				40. NTE DATE ON DETAIL			
0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9

Figure 2
Part of FOSDIC Document
for Personnel Records

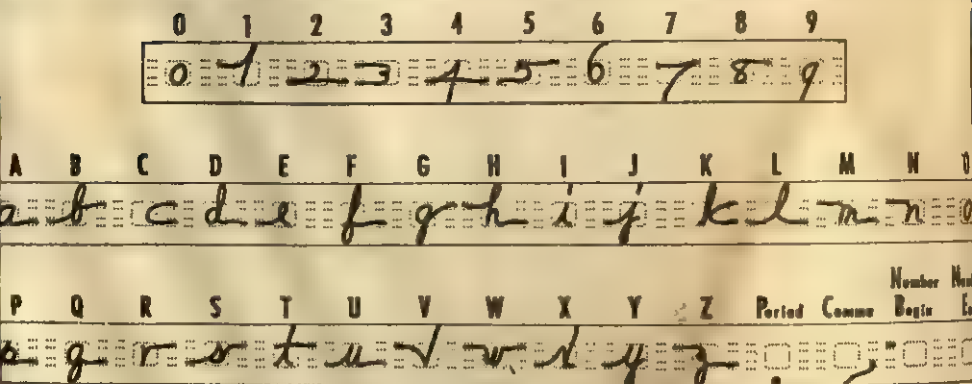


Figure 3
Alpha-numeric Characters
for FOSDIC Recognition
(Experimental)

used so far has been 16×21 inches; and the smallest, $8\frac{1}{2} \times 11$ inches.

A handy by-product of the FOSDIC system is the microfilmed records of the documents. Anyone who has ever tried to find storage space for large numbers of answer sheets will recognize the advantage of a few reels of microfilm. After the FOSDIC processing is finished, the microfilm is available for reference. Thus, the original material may be sent to dead storage or destroyed.

A possible disadvantage of the FOSDIC system is its inability at the present time to read letters. Thus, each individual must be assigned a code number, if his identity must be ascertainable from the data. The Census Bureau normally uses the Social Security number, although any unique testing number would serve as well—such as a military service number or a testing number assigned at the time of testing. This number can be recorded by the examinee, and if it is necessary to know the name, a basic deck containing the matching information can be punched.

FOSDIC's Availability

FOSDIC was available to the American Institute for Research for the Air Force Reserve Officer Training Corps project because of the nature of the organization (non-profit) and because of the nature of the data—government documents. FOSDIC as a test scoring

FOSDIC ALPHA-NUMERIC - MARKING DOCUMENT I

(GENERAL PURPOSE - 88 CHAR.)

EXAMPLE

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o

P	Q	R	S	T	U	V	W	X	Y	Z	Number Begin	Number End
p	q	r	s	t	u	v	w	x	y	z		

DATA MARKING AREA

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21																			
41																			
61																			

and analysis source will very likely continue to be available for research dealing with government data and those of similar organizations, such as school systems or institutions of higher education. Since this application is fairly new, no specific policy has yet been developed. It is not the Bureau's intent to compete with industry; but when machine time is available, personnel in the Bureau are glad to help other agencies on a reimbursable basis. The Bureau is not interested in additional work with IBM answer sheets, since use of IBM answer sheets requires the slower operation with the jig plus a deviation from the normal procedures and camera settings with which their operators are familiar. However, personnel at the Bureau are very much interested in developing general purpose answer sheets, as well as special purpose answer sheets where the volume warrants it. Eventually, it is expected that the scoring may also be done as part of the run which produces the final magnetic tape.

It is suggested that persons who foresee usefulness of the FOSDIC system as a solution to some of their test analysis or personnel records systems determine their eligibility directly from the Census Bureau, Attention: Robert F. Drury, Data Processing Systems, Bureau of the Census, Washington 25, D. C.

REFERENCE

- McNelis, David P. *Film-Optical Sensing Device for Input to Computers FOSDIC III*, Washington, D. C.: Bureau of the Census, 1961.

POLYSERIAL CORRELATION PROGRAMS IN FORTRAN

NATHAN JASPEN
New York University

THE method used in calculating biserial correlation was generalized in 1944 (Jaspen, 1946) to include triserial correlation, quadriserial correlation, and so on, where the segmented variable could be expressed in any number of categories. The assumptions are precisely the same as those made in biserial correlation: the segmented variable, although expressed categorically, is basically continuous and normally distributed. These assumptions are more likely to be true when the segmented variable is expressed in three or more categories than when it is expressed in only two categories. For instance, point biserial correlation rather than biserial correlation should be used if sex is correlated with height; and some researchers believe that item analysis, in which correlations are obtained between item scores (right or wrong) and test score, should use point biserials rather than biserials. On the other hand, if we wish to correlate course grade (expressed as A, B, C, D, or F) and test score, the assumption of continuity in the course grade can hardly be questioned. If, in addition, the assumption of normality in the distribution of course proficiency is valid, quintiserial correlation is indicated.

Normality in this instance does not mean that the letter grades are normally distributed; it means only that course proficiency is normally distributed. If the latter assumption is valid, it would not matter if 40 percent of the grades were A, 30 percent B, 20 percent C, 5 percent D, and 5 percent F. We would simply assign the upper 40 percent of the normal curve to A, and quantify the letter grade A by putting it into correspondance with the mean of the upper two-

fifths segment of the normal curve. In effect, this is what the computer is doing when it applies the formula for polyserial correlation.

The use of the segment means as class-index values representing the categories of the segmented variable permits the computation of the correlation between the segmented variable and a variable that is continuous and expressed quantitatively. The quantitative variable need not be normally distributed.

Pearson (1913) has shown that a coefficient so computed is in need of correction for "broad categories." The correction is also built into the formula.

The formula for polyserial correlation is

$$r_{psr} = \frac{\sum [(y_i - y_h)X_i]}{SD_x \sum \left[\frac{(y_i - y_h)^2}{p_i} \right]}, \quad (1)$$

in which X is the quantitative variable and Y is the segmented or categorical variable, and where

y_i = height of ordinate at lower end of interval i ,

y_h = height of ordinate at upper end of interval i ,

p_i = proportion of total group in interval i ,

X_i = mean on X of individuals in interval i ,

SD_x = standard deviation of all the X scores.

Variations of this symbolism may also be found in Peatman (1963) and in Wert, Neidt, and Ahmann (1954).

If the number of categories in the segmented variable is two, this formula reduces to the usual formula for biserial correlation.

Soper (1914) published approximations for the standard error of biserial correlation. Wert, Neidt, and Ahmann (1954) propose that the Fisher formula

$$t = \sqrt{\frac{r^2(N-2)}{1-r^2}} \quad (2)$$

be used to test the significance of polyserial r , except that they prefer to substitute for r in the t formula "the more conservative estimate from the formula for point serial correlation" (p. 272) which is identical to (1), except that the right hand term in the denominator is under a square root. The use of the square root, which eliminates the correction for broad categories, increases the estimation of

the standard error. Fisher (1950) demonstrated that formula (2) above gives better estimates of the significance of Pearson r than does the common-sense formula $t = r/\sigma_r$. The use of Fisher's z' transformation should give more satisfactory results.

Two research problems arose at about the same time that made it desirable to prepare computer programs for the computation of polyserial correlation. Both programs were written in Fortran, and both programs are operative on the IBM 7094. In addition, one of them is also operative for the IBM 1620.

In the first problem, each of the 61 items in a 61 item questionnaire permitted of 5 responses: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree. It was desired to obtain the correlation between each item and a continuous criterion. The item responses themselves were obviously continuous, and the assumption of normality (not of the distribution of responses, but of the underlying distribution of feeling tone) appeared reasonable. The computer program that was written for this situation was generalized to allow two, three, four, or five categories for each item; and it yielded the standard errors and t ratios as well as the polyserials. The number of categories for each item, which was determined by the number of categories drawing responses, could vary from item to item.

In the second problem, the reverse situation prevailed. The criterion was a trichotomous rating (High, Middle, Low), and there were nine continuous test variables. A program was prepared that computed all the intercorrelations, including the triserials with the criterion. This program was generalized to operate with criterion variables expressed in two, three, four, or five categories. The number of variables and the number of categories in the criterion variable are indicated by the program user in a parameter card, which precedes the data deck. This program has been used with both the IBM 7094 and the IBM 1620.

One virtue of computer programs is that approximations that make restrictive assumptions are unnecessary. It is true, for instance, that if the proportions in the two extreme groups are equal, the formula for triserial correlation simplifies. Jenkins (1956) proposed such a simplification, which does facilitate hand calculations and gives good results if the assumption of equal tails is reasonably correct. Burt (1944) also proposed a triserial correlation formula, in which he eliminated the middle group. The programs employed in

this paper use all the data, but do not eliminate any groups from consideration.

The only problem that arose in the preparation of the computer programs was the need to determine the ordinates of the normal curve, when a tail area is given. Hastings (1955) presented approximations which yield the ordinate when the abscissa is furnished and which yield the area when the abscissa is given, but none which yield the ordinate when the tail area is specified. Accordingly, a table of 500 ordinates was prepared at every one-tenth of one percent of the area for one side of the curve, and the table was punched into data cards. The ordinates are available to eight decimals. The programs incorporate a linear interpolation routine to determine the ordinates for areas that are four or more decimal places.

One last observation, relating to nomenclature, may be in order. Because Pearson had used the expression "biserial," the present writer used the expression "serial" as the generalization. Serial correlation has also been used, in series analysis, synonymously with "autocorrelation" (Yule and Kendall, 1950, p. 639); "intraserial correlation" (Payne and Staugas, 1955) has also been used in this connection. Therefore, the adoption of another term seems appropriate. "Multiserial," introduced by the editorial staff of *Psychometrika* (1955), is not accurate, since the correlations discussed here are zero-order correlations. Besides, it could easily be confused with "multiple serial correlation" (Wert, Neidt, and Ahmann, 1954, pp. 275-277). Therefore the term "polyserial" is proposed to signify the correlation between a categorical or segmented continuous variable and a quantitative variable.

REFERENCES

- Burt, C. "Statistical Problems in the Evaluation of Army Tests." *Psychometrika*, IX (1944), 219-235.
- Fisher, R. A. *Statistical Methods for Research Workers*. New York: Hafner, 1950, pp. 193-204.
- Hastings, C. *Approximations for Digital Computers*. Princeton: Princeton University Press, 1955.
- Jaspens, N. "Serial Correlation." *Psychometrika*, XI (1946), 23-30.
- Jenkins, W. L. "Triserial R—A Neglected Statistic." *Journal of Applied Psychology*, XL (1956), 63-64.
- Payne, M. C. and Staugas, L. "An IBM Method for Computing Intraserial Correlations." *Psychometrika*, XX (1955), 87-92.
- Pearson, K. "On the Measurement of the Influence of 'Broad Categories' on Correlation." *Biometrika*, VII (1913), 96-105.

- Peatman, J. G. *Introduction to Applied Statistics*. New York: Harper, 1963, pp. 128-130.
- Psychometrika*, XX (1955), "Subject Index," 338-350.
- Soper, H. E. "On the Probable Error of the Biserial Expression for the Correlation Coefficient." *Biometrika*, IX (1914), 91-115.
- Wert, J. E., Neidt, C. O., and Ahmann, J. S. *Statistical Methods in Educational and Psychological Measurement*. New York: Appleton-Century-Crofts, 1954, pp. 271-275.
- Yule, G. V. and Kendall, M. G. *An Introduction to the Theory of Statistics*. New York: Hafner, 1950.



A 1620 FORTRAN PROGRAM FOR COMPILING A FLANDERS-AMIDON INTERACTION ANALYSIS MATRIX

ROBERT M. RIPPEY
University of Chicago

A number of methods have been suggested for making reliable observations of the interaction between students and the teacher in the classroom. Medley and Mitzel presented a comprehensive survey of a large number of such observational techniques in *The Handbook of Research on Teaching*, edited by Gage (1963). In comparing these many systems, Medley and Mitzel (1963) stated: "Flanders . . . has developed the most sophisticated technique for observing climate thus far, one that is unique in that it preserves a certain amount of information regarding the sequence of behavior" (p. 271).

Although Medley and Mitzel gave a satisfactory condensation of the system, further information regarding the details of the method, and applications to studies of classroom behavior may be found in publications by Amidon and Flanders (1962).

Interaction analysis consists of a scheme for the classification of classroom verbal behavior into ten categories. Observations are recorded at three-second intervals. These observations consist of numbers from one to ten which correspond to each of the ten categories of verbal behavior. Classroom activities are divided into episodes by means of double lines drawn through each set of observations. When the observations are completed, a separate matrix is assembled for each episode. The 10×10 matrix contains 100 cells, each cell standing for a pair of contiguous verbal behaviors. For example, if an instructor asked a question, and if this question were followed by a student response, the observer would record a four, followed by an eight. These two numbers stand, respectively,

for the verbal behaviors of teacher questions and student responses. This pair of sequential observations would further be recorded as a single tally in the (4, 8) cell of the interaction matrix. Upon completion of the matrix, columnar totals are calculated, and certain ratios representing the distribution of verbal behavior are calculated. In using interaction analysis with teachers for the purpose of giving them objective reports of their classroom verbal behavior, as well as in employing interaction analysis in experimental classes for research, the author found the compilation of a matrix for a single episode extremely time-consuming. Therefore, a computer program was written which takes approximately 30 seconds on the IBM 1620 computer for preparing a matrix and for making the necessary calculations for up to 1000 observations.

Computer Input

Observations are placed directly on punched cards. Single digits corresponding to the observed verbal behavior categories are punched in sequence. As many cards as are necessary may be used to record the observations provided that the total number of observations does not exceed 1000. Any number of episodes up to 999 may be made in one loading of the program. Each episode must be preceded by a header card containing a) the number of fully punched cards in the first two columns and b) the number of columns used in the last card of the episode punched into columns three and four. In the preparation of the header card, it is important to make certain not only that the last card of each episode is not counted in columns one and two, but also that the number of columns punched in the last card is always indicated in columns three and four. For example, if the last card of an episode contains 80 punches, one does not count this card in the first two columns of the header card, but enters the number 80 in columns three and four of the second header card. One slight modification of the verbal categories is necessary. Verbal category 10 must be punched as a 0, although it will be interpreted as a 10 in the output.

The first data card must be a control card with a three digit number punched in columns one, two, and three. This number indicates the total number of jobs to be run. If a single job is to be run, the first card must have 001 punched in it. The program automatically numbers each matrix in sequence. It is desirable to have a list identifying each matrix by number prior to running.

TABLE 1

Interaction Analysis Matrix of Classroom Observations

	1	2	3	4	5	6	7	8	9	10	Sum
1	0	0	0	0	0	0	0	0	1	1	2
0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	15	0	1	0	0	0	2	1	19
0	0	0	0	0	0	0	0	5	2	0	7
5	1	0	0	0	197	0	0	0	12	5	221
6	0	0	0	0	0	3	0	0	1	2	6
7	0	0	0	0	2	0	0	0	0	0	2
8	0	0	0	0	5	0	0	0	0	0	5
9	1	0	4	1	13	3	2	0	11	0	35
10	0	0	0	0	3	0	0	0	6	21	30
Total	2	0	19	7	221	6	2	5	35	30	327

Teacher Student Talk Ratio: 6.4250.

Indirect Direct Ratio: .1222.

Revised Indirect-Direct Ratio: 2.6250.

Content Ratio: 1.3944.

Indirect Direct Ratio (89): .2808.

Steady State Ratio: .7553.

Computer Output

The matrix output of this program is punched. When the punched output is printed, it resembles the matrix shown in Table 1.

Using this program, one finds that it is possible to compile large numbers of observations in less time than it would take to prepare a single matrix by hand. Furthermore, errors rarely occur.

Occasionally an error message will be punched along with one of the ratios. This event indicates that certain verbal behaviors were lacking in the episode, and that, as a result, the program had attempted division by zero. This difficulty could have been eliminated from the program, but was not, for the error message serves as a useful signal that certain verbal behaviors were not present in the episode.

A printout of this program is available from the author on request.

REFERENCES

- Amidon, E. and Flanders, N. *The Role of the Teacher in the Classroom*. Philadelphia: Group Dynamics Center, Temple University, 1962.
- Medley, D. H. and Mitzel, H. E. "Measuring Classroom Behavior by Systematic Observation." In N. L. Gage (editor), *Handbook of Research on Teaching*. Chicago: Rand McNally and Company, 1963, pp. 247-326.



COMPUTER SEARCH FOR GROUP DIFFERENCES

M. CLEMENS JOHNSON
The University of Michigan

THIS paper describes a new procedure for processing data in problems aimed at uncovering differences between two groups of individuals. In the procedure a digital computer is used to combine qualitative characteristics and to determine the extent to which each combination is present within the two groups. It is assumed that the possible number of combinations is very large. No assumption is made concerning the form of the variable distributions or intercorrelations of measurements.

While data processing is carried out by digital computer, the underlying rationale is simple. The goal is to identify combinations of qualitative characteristics which are more effective in separating the two groups. With its great speed the computer can track down more promising combinations among the many possibilities.

These combinations can be investigated more intensively, and with other samples. For instance, the results of the computer search could suggest characteristics for use in expectancy tables, or factors to be controlled in experimental studies. In effect, the computer may serve as a hypothesis generator.

Some results of applying the approach to a particular set of data for 2 samples of individuals are reported. The example involved analysis of questionnaire responses provided by a sample of high school students. The objective was to identify combinations of responses which indicated greater divergence of opinion among boys and girls.

For purposes of clarity subsequent discussion is divided into two sections: (a) the computer search procedure, and (b) some results of applying the computer search.

Computer search procedure. Computer search appears applicable to a wide range of classification problems involving two groups of individuals. Examples may be found in studies of characteristics of students who drop out of school and students who are graduates; persons with disease x and persons without disease x ; persons who vote for Republican candidates and persons who vote for Democratic candidates.

The approach is suited to classification problems in which the majority of measurements are qualitative rather than continuous. Ratings, diagnoses, and other judgments applied to individuals are examples of qualitative data. When data are qualitative each measurement takes only a finite and usually small number of values.

A specific computer program has been written to process data in the manner described. In an application of this program the data pool can include 75 factors per individual, each factor being defined at either 2 levels or 3 levels. Examples of factors at 2 levels are: problems scored as pass or fail; tasks rated as easy or difficult; questions answered as yes or no. A value for a continuous variable may be classed as above or below the median. Examples of factors at 3 levels are: attitudes expressed as agree, undecided, disagree; rural, urban, suburban; and so on. Thus the maximum number of factor-level characteristics is $75 \times 3 = 225$. The maximum number of individuals in the combined groups is 300.

Because rapid data processing requires that all information for all individuals be available in core storage, a computer with large memory is required. The particular program was written for the IBM 7090 data processing system using the MAD language (Michigan Algorithm Decoder).

In the operation of this program the computer searches for pairs of factor-level characteristics and later for combinations of three. Arbitrary limits are placed upon the number of times it will search the data. Selection of combinations is at random, although systematic search of all possibilities could be accomplished in many instances.

While searching for better discriminators among the combinations, the computer assumes no knowledge concerning the extent to which individual characteristics occur within the two groups. Repeatedly, a combination of characteristics is selected at random and proportional occurrence is determined for each group. The only in-

formation carried over as the computer searches one combination after another is the value for the largest difference in proportional occurrence among the two groups.

When the computer finds a difference in proportional occurrence greater than that observed for any previous combination, a print-out provides (a) the search number at which the combination was observed, and (b) frequencies and proportional occurrence within each of the two groups. While continuing to search for a combination with still higher discriminating power, the computer will also print as output combinations which are in the vicinity of the tentative maximum.

The program allows the user to hypothesize particular combinations to be checked by the computer. The effectiveness of these combinations can then be compared with those identified by the computer through the random search procedure.

Some results of applying computer search. A social studies questionnaire¹ was given to 175 students in five Michigan high schools. Students were in grades 10, 11, and 12. From the pooled samples a random selection was made of 74 boys and 74 girls. Responses of these students provided data used in the analysis.

The questionnaire, *A Survey of High School Opinion*, requires that a student give an opinion toward each of 30 statements. The statements generally refer to the operation of democratic processes. Examples are, "Juries in our community are chosen fairly." "Taxes are too high in our community."

In the first computer run, tabulations were obtained for each response to each statement. Responses were coded as agree, undecided, or disagree so that the total number of responses was $3 \times 30 = 90$. Tabulations revealed differences in proportions falling in the agree, undecided, and disagree categories within each sex. An example is provided by the statement, "The welfare state tends to destroy individual initiative." Among the boys, 50 percent agreed, 33 percent were undecided, and 17 percent disagreed. Responses of boys and girls appeared more similar than different. There was no single response, for instance, which differentiated a

¹ The questionnaire was developed and compiled by S. E. Dimond, Professor of Education, University of Michigan. Included in the questionnaire were statements from H. H. Remmers and D. H. Randler, *The American Teenager*, (Boobs-Merrill, 1957), and P. E. Jacobs, *Changing Values in College*, (Harper, 1957).

large majority of boys from girls. Responses which produced more divergence of opinion between the sexes were the following.

"The Communist Party should be outlawed." Disagreeing with this statement were 39 percent of the boys, but only 16 percent of the girls.

"Politics is a messy business." Undecided were 38 percent of the boys and 18 percent of the girls.

While the first computer run produced only frequency counts for single responses, the second run involved computer search for pairs of responses. About 4,000 random pairs (with some duplication) were checked by the computer. Output data consisted of pairs of responses identified as more effective than others in separating the opinions of the 74 boys and the 74 girls.

It was found in this example that the search for pairs of responses produced results similar to those for single responses. No one pair differentiated a large proportion of boys from girls, or conversely. The computer was able to identify pairs which characterized one-fourth to one-third of one sex, and few of the other. These pairs were:

- (1) "Most students would cheat on an exam if they were sure of not being caught."
- (2) "If men are not crooked when they go into politics, they are crooked when they come out."

Of the girls, 26 percent agreed with (1) and were undecided about (2). Only 4 percent of the boys responded in the same way.

- (1) "Newspapers and magazines should be allowed to print anything they want except military secrets."
- (2) "Generally speaking, Negroes are ambitious and intelligent."

Of the girls, 38 percent disagreed with (1) and agreed with (2). Of the boys, only 8 percent.

- (1) "A large mass of the people are not capable of determining what is and what is not good for them."
- (2) "Some of the petitions which have been circulated should not be allowed by the government."

Among the boys, 31 percent agreed with (1) and disagreed with (2). Of the girls, only 7 percent gave the same responses.

While one pair of responses did not differentiate most of the boys from the girls, it is possible that such a goal could have been realized through multiple pairs. For instance, it might be expected that

a boy could be characterized by at least one among several alternative pairs of responses, while a girl would be characterized by none. The analysis was not made in this study.

In a third run the computer searched for combinations of three responses which were different from the group of boys and of girls. More than 100,000 combinations ($C_3^{30} \times 3^3$) were possible. Random selection was employed by the computer in checking about 18,000 of the possibilities.

Many "better" triples were identified and printed as output. However, no one combination of three responses was superior to the best pair in discriminating the boys and girls. An example of a more discriminating triple is shown.

- (1) "Labor organizers should be excluded from a town where there has been no labor trouble."
- (2) "A large mass of people are not capable of determining what is and what is not good for them."
- (3) "Juries in our community are chosen fairly."

Of the boys, 18 percent agreed with all three statements. Of the girls, only 3 percent. Some mental concentration is required to evaluate combinations of three responses. The difficulty is emphasized in this application where each statement was coded to permit three possible responses. (Interpretation would tend to be facilitated in an application where the various characteristics were either present or absent.)

Because the program is new and relatively untried, further information is not available on operating characteristics. Information is lacking also on the mathematics underlying the data processing.



BOOK REVIEWS

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

<i>Bloom's Stability and Change in Human Characteristics.</i>	247
LEWIS R. AIKEN, JR.	
<i>Adams' Measurement and Evaluation in Education, Psychology, and Guidance.</i> JOAN J. BJELKE	250
<i>Stanley's Measurement in Today's Schools.</i> ALLAN R. STARRY	252
<i>Barnette's Readings in Psychological Tests and Measurements.</i>	254
STEPHEN W. BROWN	
<i>Bauernfeind's Building a School Testing Program.</i> DORIS S. BARTLETT	255
<i>Sanford's The American College: A Psychological and Social Interpretation of the Higher Learning.</i> C. RUSSELL DE BURLO, JR.	257
<i>Festinger's Conflict, Decision and Dissonance.</i> JOHN DELA-MATER	259
<i>Mouly's Science of Educational Research.</i> GENE V. GLASS AND DALE E. MATTSON	259
<i>Havighurst, Bowman, Liddle, Matthews, and Pierce's Growing Up in River City.</i> ALMA V. WILLIAMS	263
<i>Arons and May's Television and Human Behavior.</i> V. C. ARNSPIGER	265
<i>Ammons and Ammons' The Quick Test (QT): Provisional Manual.</i> PETER F. MERENDA	268
<i>Murstein's Theory and Research in Projective Techniques (Emphasizing the TAT).</i> FLORENCE DIAMOND	271
<i>Jenkins' The Morgan State College Program—An Adventure in Higher Education.</i> JULIAN C. STANLEY	273

<i>Ohlsen's Guidance Services in the Modern School.</i> MABEL E. HAYES	276
<i>Garry's Guidance Techniques for Elementary Teachers.</i> ROY M. FITCH	277
<i>Bennis, Schein, Berlew, and Steele's Interpersonal Dynamics: Essays and Readings on Human Interaction.</i> ARTHUR LERNER	278
<i>Foss' Determinants of Infant Behaviour II.</i> JEROME E. LEAVITT	279
<i>Garrison, Kingston, and McDonald's Educational Psychology.</i> PHILIP S. VERY	280
<i>Drayer's Problems and Methods in High School Teaching.</i> REGINALD L. JONES AND L. WARREN NELSON	281
<i>Deutsch's Selected Papers from the Institute for Developmental Studies [of the New York Medical College], Arden House Conference on Pre-School Enrichment of Socially Disadvantaged Children.</i> JULIAN C. STANLEY	282
<i>Roswell and Natchez's Reading Disability: Diagnosis and Treatment.</i> WALTER PAUK	286
<i>Leedy's Read with Speed and Precision.</i> WALTER PAUK	287
<i>Engle's Psychology.</i> ROY M. FITCH	288

Stability and Change in Human Characteristics by Benjamin S. Bloom. New York: John Wiley & Sons, Inc., 1964. Pp. xiv + 237. \$7.00.

This book, which was begun when the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, attempts to make sense out of the available longitudinal data on development.

In the preface, the author summarizes the import of the book in three different ways. One way is a descriptive equation, $I_2 = I_1 + f(E_{2-1})$, where I_1 and I_2 are measures of a characteristic at points 1 and 2 in time and E refers to environmental conditions affecting I in the interval between the two I measures. A more specific way of summarizing the book is by plotting level of development as a negatively accelerated function of age, a graph which also indicates that the limits of variation to the function which environment can produce decreases with age. The third way of summarizing the book is by the following propositions.

1. The relation between parallel measurements over time is a function of the levels of development represented at the different times.
2. Change measurements are unrelated to initial measurements but they are highly related to the relevant environmental conditions in which the individuals have lived during the change period.
3. Variations in the environment have greatest quantitative effect on a characteristic at its most rapid period of change and least effect on the characteristic during the least rapid period of change." (p. vii)

The text of the book consists of seven chapters in which these ideas are delineated and elaborated through the medium of examples of various physical and psychological characteristics.

Chapter 1 discusses methods of studying stability and change and the major longitudinal studies which have been conducted. Such questions as the degree of similarity between the results of different longitudinal studies, the congruence between cross-sectional and longitudinal data, and the magnitude of the effect of extreme en-

vironments and early experience are presented as issues to be dealt with in the book.

Chapter 2 considers the correlational data from a dozen longitudinal studies of height as a model for the analysis of other types of growth. Data on weight and strength changes with age are also considered briefly. Three versions of the "Overlap Hypothesis," which attempts to predict the degree of overlap between two sets of measurements in a longitudinal series, are presented. Anderson's version of this hypothesis, which suggests that "... for longitudinal data the correlation between the two measurements would equal the square root of the ratio of the two means $\sqrt{M_{x1}/M_{x2}}$ if the relationship between the initial scores and final scores was approximately zero" (pp. 27-28), however, is the one applied most frequently to data presented in the book.

Chapter 3 is devoted to the analysis of the changes in scores on intelligence tests. It is concluded that when, for the various studies, the correlation between intelligence at each age and intelligence at maturity is plotted against chronological age, the shapes of the curves approximate that of Thorndike's absolute scale of intelligence development.

Chapter 4 deals with the stability of achievement test data. From an analysis of the incomplete data available, the author concludes that there is great similarity of findings in different longitudinal studies of achievement. The relationships of teachers' marks, scores on achievement tests, and reading comprehension over time are similar for these three indices of achievement and are approximated by values inferred from an absolute scale based on normative data on vocabulary development.

Chapter 5 is concerned with analyzing the data on the longitudinal development of interests, attitudes, and personality, which the author believes are far from satisfactory. Although the "Overlap Hypothesis" in relation to an absolute scale of development was not found applicable to these data, the author was able to make some statements about the percentages of the "adolescent variance" in certain traits present at earlier ages. This analysis indicates that the longitudinal development of interests, attitudes, and personality is quite different from that of physical characteristics such as height. The author concludes that the development of the former continues throughout life and approximates an age curve (equal growth units per year) of development.

Chapter 6 considers the effects of environment on both stability and change and states three major propositions about longitudinal measurement.

1. The correlation between measurements of the same characteristic at two different times is a function of the Overlap Hy-

- pothesis when the environment in which the subjects lived during the intervening period is *not* known or considered.
2. The correlation between measurements of the same characteristic at two different times will approach unity when the environment in which the individuals have lived during the intervening period is known and taken into consideration.
 3. The gains made by individuals subjected to the same powerful environment will tend to be equal." (Pp. 199-200)

Chapter 7 presents some conclusions and implications of the propositions and data presented in Chapters 1-6. In addition to reiterating his hypotheses and findings, the author expresses his hope for new longitudinal studies and improved methods of measuring individual characteristics and environments. Finally, certain major issues such as "maximization versus optimal growth and development," "ideal environment," and "social responsibility" are aired and related to the theses of the book.

There is much in this book which should be brought to the attention of the specialist in the area of development as well as any educated person with an interest in the field. Attempts to find general principles in the behavioral sciences are so often met with frustration that it is heartening to find even a partial success. In the opinion of the reviewer, that is what is represented by this book.

Although the findings indicate that the "Overlap Hypothesis" may approximate the empirical results for some of the characteristics studied, it is far from being a general law of development. Many psychological data have appeared initially to fit the almost classic negatively accelerated curve, but further study of the domain has shown that there are many variables which affect the shapes of the curves obtained. In all fairness, the author admits in various places throughout the book that his hypotheses are tentative ones and that different methods of measurement, for example, may change the outcomes somewhat. However, one has to search for these "qualifications." For the most part, the reviewer obtained the impression that these hypotheses are the author's "credo" and that he was trying to prove them rather than test them.

The author is not always explicit about his method of data analysis, for example, on pages 100-101. In addition, it is not always clear what the author means by a particular concept, such as that of intelligence in Chapter 3. Is it test score, mental age, intelligence quotient, an intervening variable, or what?

Although, in the main, the author has a readable style, he tends to overgeneralize and to be too speculative in spots, and the book is too repetitious. It is obvious that much effort went into the production of this book, but perhaps it would have saved some time if the author had been more succinct. For example, large portions of

Chapters 6 and 7 could easily have been omitted with no loss to the meaning of the book. In the reviewer's opinion, this material should have been published as a monograph rather than in book form.

LEWIS R. AIKEN, JR.

University of North Carolina
at Greensboro

Measurement and Evaluation in Education, Psychology, and Guidance by Georgia Sachs Adams in consultation with Theodore L. Torgerson. New York: Holt, Rinehart and Winston, 1964. Pp. xiii + 654.

The increasing depth of instruction in the field of measurement and evaluation in our teacher training institutions is reflected in the volume *Measurement and Evaluation in Education, Psychology, and Guidance*. Realistically emphasizing the complexities of educational measurement, the text provides a rigorous and comprehensive presentation of basic evaluation techniques for (1) the classroom teacher, (2) the school psychologist or psychometrist, and (3) the administrator.

Comprised of 17 chapters, the volume is structured around 4 major divisions: "Basic Principles and Procedures" (Part I), "The Study of Individuals" (Part II), "The Improvement of Instruction" (Part III), and "Administrative, Supervisory, and Guidance Aspects of Measurement and Evaluation" (Part IV). The volume departs from the traditional format of confining the statistics to a separate section. Rather, there is an integration of the conceptual and applied aspects of the various constructs, the primary emphasis being on the conceptual. Lucid and complete tables are liberally supplied in the technical chapters to reinforce major points covered in the text. Each chapter also includes numerous additional references for further reading.

In Part I the tone of the text is set. While developing the relative merits of the evaluative process, the text carefully delineates the limitations and difficulties encountered. Following a chapter on types of test scores and norms are comprehensive and insightful chapters on reliability and validity. Cogent explanations are offered of the various methods of computing reliability and the concomitant sources of variance. Included also are topics such as interpretation of r^2 in terms of percent of variance accounted for, reliability of difference scores, factors affecting the size of reliability coefficients, and means of improving the reliability of test scores. Equally as complete and rigorous is the chapter on validity. Supplementing the text are clear, concise, tables presenting examples and procedures for the development of tests designed to measure each type of validity. Discussion questions are provided with each of these chapters to assist the student in applying the concepts. While the ques-

tions following the validity chapter are very complete, the reviewer felt that it would have been helpful to the student to have more questions requiring practical applications of reliability, for understanding principles of validity and reliability is basic to comprehending the remainder of the text. Part I concludes by applying the concepts of norms, reliability, and validity to the problem of selecting tests.

In Part II attention is focused on actual measurement of the individual. Included are separate chapters covering the measurement of aptitudes, interests and attitudes, personal and social adjustment, and personality inventories and projective techniques. The student is asked to apply the knowledge gained in Part I on interpretation of test scores to achieve a more thorough understanding of the relative merits of the various tests. Both the strong and weak points of each type of test are realistically evaluated to assist the student in interpreting scores.

Rather than superficially discussing several tests, each chapter explains in considerable detail a few major tests, such as the *Binet* and *WISC*. Supplementing this discussion is an exhaustive Appendix of classified, published tests. The reviewer felt that this method of presentation was extremely effective, because it offers the student an insight into the demanding requirements of actual test analysis. In this manner the volume stresses the importance of consulting the experts before selecting and using any particular test. The discussion questions, especially those following Chapter 6, "Measurement of Aptitudes," provide the student an opportunity to make practical applications of the theory discussed.

In Part III the heavy weighting given to teacher-made tests reflects the importance with which the author views the role of interaction of evaluation and instruction. The author's statement, "Teacher-made tests have great potentialities for enriching or limiting the students' self-directed study," is one which cannot be repeated too often. Many excellent examples of the various types of test items are provided along with a highly useful "Checklist for Discovering Faulty Test Items and Other Errors in the Construction of Objective Tests." Supplementing the textual materials, tables present succinct illustrations of the relative advantages and disadvantages of essay and objective test items. In the chapter on achievement tests, again only selected tests, such as *STEP* and the *SRA* series are discussed in detail, the student being referred to the Appendix of published tests in various fields. Whereas most basic measurement texts merely refer to Bloom's *Taxonomy of Educational Objectives*, this volume devotes an entire chapter to a discussion of 20 subheadings of the *Taxonomy*. Another special feature of this volume is Chapter 12, "Evaluating Student Performance in the Skills," offering the educator with needs in this area a rare op-

portunity to see measurement theory actually applied to his field. Concluding Part III is an excellent chapter relating the major concepts to educational diagnosis, both group and individual. The discussion is supplemented with visual examples of diagnostic situations.

In Part IV the role of administrators, supervisors, and guidance personnel in evaluation is discussed. Although less attention is devoted to this area, actual practical application of test data in both group and individual guidance is presented. Thus a final reinforcement of the previous conceptual discussions is provided.

The real strength of this volume lies in its readable and comprehensive coverage of the many phases of measurement and evaluation and in its strong emphasis upon the importance of the role of the teacher in this area. The many footnotes and lucid tables and appendices offer the serious student a rigorous treatment of subject matter. Pedagogically, this volume would provide an excellent basic text for a test and measurement course. However, in spite of the author's statement that the text had been designed for the beginner, the reviewer felt that the student should have a background in basic statistics to derive maximum benefit from this highly conceptual treatment. Chapters such as that on converted test scores could prove very threatening to the statistically unsophisticated student. The reviewer does not offer this opinion as a criticism, but rather as a limitation of the volume, because it is felt that this type of presentation adds further impetus to the upgrading of tests and measurement courses.

JOAN J. BJELKE

University of Southern California

Measurement in Today's Schools (Fourth Edition) by Julian C. Stanley. Englewood Cliffs, N. J.: Prentice-Hall, 1964. Pp. xviii + 414. \$7.50.

This edition marks the second revision undertaken by Stanley of the popular text introduced by C. C. Ross in 1941. The entire book has been substantially reworked and, as the author puts it, made "less folksy." Missing from the current volume are the "Measurement in Instruction" section and many of the long quotations. The effect of these deletions on the overall length of the book has been minimized by enlarging the remaining 11 chapters and adding four new appendices.

All too frequently future teachers are confronted with only one course or with segments of one measurement course which must somehow provide them with the skills required to develop adequate testing programs and render sound measurement decisions. To further complicate the measurement instructor's problem, a significant percentage of these students has neither the background nor the

motivation to assimilate this essential material within the brief span of a single semester. Stanley appears to have tailored his volume for just this type of audience and learning situation. He assumes no specific preparation on the part of the reader and attempts to stimulate interest by emphasizing the applications of basic measurement and evaluation principles to everyday teaching experiences. Algebraic and other statistical symbolism is avoided; it is replaced whenever possible with a systematic "talking through" of concepts for the non-mathematically inclined. Illustrations and examples, utilized generously throughout, are borrowed from current standardized tests and other literature with which the teacher may come into direct contact. In general, a realistic effort is made to increase the student's awareness of a wide variety of measurement topics and issues instead of seeking to achieve a high degree of comprehension in a few areas.

The author states that the book falls into two main parts, with the first five chapters devoted to fundamental principles and the last six chapters to the construction and use of measuring instruments. Chapter Two is of particular interest, as it furnishes a thoroughly researched historical sketch of educational and psychological measurement which should prove to be fascinating reading for even the most advanced student.

Several critical topics which are poorly understood by many teachers and superficially treated in some similar textbooks are extensively discussed. Among these are the concept of reliability, the various types of measurement errors, the relative advantages and disadvantages of different types of tests, the effects of correction formulas, and the importance of guessing instructions. Although the section concerning guessing tends to be a trifle confusing, it is none the less informative and correctly reveals the lack of agreement among writers on this subject.

Stanley's realistic approach is exemplified in his presentation of standard scores. Not only are the customary z -transformations difficult for the average student to grasp quickly, but also it is doubtful that teachers will take the time to compute them routinely for their own tests. With perhaps these and other considerations in mind, the author has elected merely to describe z -scores briefly and to concentrate on the construction, interpretation, and application of normalized stanine and sta-eleven scores. Tables are provided which reduce the computations involved to a convenient counting operation. The net result is a teachable standardization technique with sufficient accuracy and simplicity for everyday use by the teacher.

The ten appendixes provide short-cut computational techniques and practice exercises. Item difficulty and discrimination indices, which receive cursory treatment in the body of the book, are given further attention in Appendix A, "A Simplified Test-Analysis Pro-

cedure." Some instructors are likely to feel that these statistics deserve a more prominent position in such a text.

Conspicuous by its absence is the term "aptitude" as it is commonly used (or misused) to denote a particular category or class of tests. Although this is undoubtedly an intentional omission, the popularity of the term among educators is sufficient in itself to justify a detailed examination. It might be argued on somewhat similar grounds that the introduction of a correlational method of item analysis would better prepare the student to understand the associated terminology when it appears in some future situation.

In summary, this is a praiseworthy volume with no major shortcomings. It can be highly recommended in a teachers' self-improvement program or as the basic textbook for elementary measurement and evaluation courses.

ALLAN R. STARRY

*Measurement and Research Center
Purdue University*

Readings in Psychological Tests and Measurements by W. Leslie Barnette, Jr. (Editor). Homewood, Ill.: The Dorsey Press, Inc., 1964. Pp. 354 + xi.

The desirability of assigning journal articles as "outside reading" for undergraduate courses is frequently offset by two major disadvantages: (1) the supply (library or otherwise) is usually far below student demand and (2) journal materials often contain nomenclature and/or procedures beyond the command of the average undergraduate student. Faced by this dilemma, Dr. Barnette, of the State University of New York at Buffalo, selected, categorized, and edited this softbound anthology of 48 articles concerning psychological measurement.

The 48 articles represent contributions from many of the well known experts (e.g., Guilford, Thurstone, French, Kuder, Strong, Flanagan, Cattell, and others), as well as articles by some of the lesser known researchers and theoreticians. In the reviewer's opinion these selections show a good balance between recent advances and some of the older more classical literature. The student may, therefore, gain "first-hand" familiarity with both types of material.

The book is divided into 11 sections, as follows: (1) General Measurement Problems (4 articles); (2) Test Administration Problems (4 articles); (3) Norms (2 articles); (4) Response Set (3 articles); (5) Reliability (2 articles); (6) Factor Analysis (3 articles); (7) Validity (17 articles); (8) Intelligence (2 articles); (9) Personality (5 articles); (10) Interest (3 articles); and (11) Critiques of Testing (3 articles). For the most part the articles are comprehensive in their coverage of the topics. One notable exception to this is section 8 concerning intelligence. The two articles, "The Mo-

tivation Factor in Testing Supervisors" by E. E. Jennings, and "A Study of Individuals Committed to a State Home for the Retarded Who Were Later Released as Not Mentally Defective" by S. L. Garfield and D. C. Affleck, do not seem to represent either major areas of current research or classical literature in the field. However, since other sections contain articles on Wechsler's contributions to the measurements of intelligence, Guilford's structure of intellect, and Thurstone's primary mental abilities, the topic of intelligence appears to be adequately covered.

The reviewer would like to point out some of the exceptionally good features of this anthology: (1) there is a single comprehensive bibliography, (2) there are both subject and author indices, (3) some of the more humorous articles, e.g., "Validity, Reliability and Baloney," by Cureton, are included, as well as the more serious literature, (4) portions of the criticism by Whyte, Hoffmann, and Shlien are presented, which serves to let the student know what the student may face among individuals who are antagonistic toward testing, (5) there are excellent editorial comments preceding each article, and (6) there are relatively few errors.

In editing this volume Professor Barnette sought to eliminate much, if not all, of the statistical aspects of particular articles. Although this procedure may be easily defended in terms of teaching the large majority of students in introductory courses, in the reviewer's opinion this lack of consideration of statistics detracts from the usefulness of the book as a supplement for students planning graduate study in test construction or psychometrics. As such, the book is not recommended for this audience. However, it is felt that this volume will prove very useful to, and will be well accepted by, most professors and students involved in introductory measurement and industrial psychology courses.

STEPHEN W. BROWN
University of Southern California

Building a School Testing Program by Robert H. Bauernfeind. Boston: Houghton Mifflin Co., 1963. Pp. 343.

This is a textbook prepared for instructors in educational and psychological measurement and evaluation, graduate students, testing directors, and school administrators. Since one of the author's purposes in writing this text was to make available a presentation and discussion of such instruments as are specific to school use, he has limited his consideration to paper-and-pencil group tests.

After presenting a brief history of the growth of large-scale school testing in this country, the author plunged into a discussion and explanation of statistical techniques and concepts essential for making sensible choices of standardized tests. He elaborated on statistical concepts for research purposes appropriate to the school

administrator or testing director. In the bulk of the book the reader is led adroitly through a maze of achievement tests from kindergarten through high school. In the final chapters are an assessment of the paper-and-pencil techniques for vocational interest and personality measurement, a consideration of the growth of subject-matter tests, suggested approaches to building a school testing program, and a look at the future of large scale school testing.

The content of this book is probably best conveyed by Bauerfeind's frankly stated opinions which are as follows: (a) achievement tests are the "heart" of a school testing program, (b) tests of mental ability are of questionable value in measuring student development, (c) standardized testing in the early primary grades also seems a dubious practice, (d) personality and interest measurement have not proved too useful for predictive purposes, and (e) forced-choice inventories are useless except to predict forced-choice behavior. He does favor the following practices: (a) the development of local norms and stanines used in relation to published national norms, (b) employment of teacher ratings for personality evaluation, (c) a cautious use of free-response inventories, and (d) mental ability tests based on pictorial materials. He is hopeful that multi-factor aptitude tests will eventually prove more helpful as the criterion situations are refined or changed. The emphasis is on the use of batteries that not only test in series but also provide opportunity for evaluation of development. This is a partial indication of the opinions and biases presented. For the most part they are thoroughly documented and well-argued throughout the book. Many tests are not only presented in detail but also discussed in relation to the practical needs of school personnel who must make the choices.

For the administrators and testing directors this book should be very helpful. It is a clear presentation of necessary information for those who are too frequently faced with a confusing morass of test brochures, catalogues, and pamphlets that pile in from Maine to California. The directors of testing programs are aided in creating a process for selecting and using tests in schools.

As a textbook in a course in educational and psychological evaluation it can only be supplementary to the established text. The scope of the book is admittedly limited, with little mention of either diagnostic tests or consideration of sociometric techniques that have proved valuable in school group situations. There is also a lack of attention to the role of individual testing in the school testing program. The instructor's guide pamphlet seems extraneous and redundant, since chapter summaries, suggested questions, and assignments are already given in the body of the text.

The point of view expressed in this book affirms the experience of many individuals who do testing or who teach in the field of edu-

cational measurement—particularly in the expressed hope that as the number of individual instruction programs increases large-scale school testing will decrease. The strongest criticism of the book would pertain to its omission of certain topics. The discussion of the preparation of teachers for understanding the tests that they administer and frequently interpret is inadequate. It is puzzling that the discussion of "culture-free" tests was so cursory when as a result of school integration efforts there is an unprecedented need for instruments of evaluation for placement and planning curricula. It does seem appropriate at least to mention the challenge of this social upheaval to those involved in "building school testing programs."

DORIS S. BARTLETT

Bank Street College for Teachers

The American College: A Psychological and Social Interpretation of the Higher Learning by Nevitt Sanford (Editor). New York: John Wiley & Sons, 1962. Pp. xvi + 1084.

The editor has employed the contributions of 30 individuals, including himself, primarily from the disciplines of psychology, sociology, or anthropology to form a monumental work of 29 chapters dealing with the undergraduate liberal arts college. The chapters are organized into eight functional areas around, within, and through the individual be he student or teacher or, rarely, the administrator.

The opening chapters analyze the student by "... attempting to summarize what is known and to indicate what might be known about the characteristics, susceptibilities, and potentialities of those who enter college." Then follows the student's environment, the factors within the college that play on the student—factors "... which are chiefly the educational activities of teachers, that might be presumed to influence him." Following the environmental aspect of the college, the chapters discuss the behavior of college students "... on patterns of interaction between students and particular processes of the college." Finally, the analysis of the undergraduate liberal arts college turns to the "... effects upon students of the college experience."

Professor Nevitt Sanford, the editor and one of the contributors, states that "... the major purpose of this volume is to help put the resources of the newer social sciences into the service of liberal education." Indeed, the central thesis of this symposium is that educational change and curricular innovation have as a main barrier the lack of a scientific basis for educational practice. Therefore, a scientific approach to higher education is a vital need, and the science to be employed to eliminate the barrier to educational change lies in general psychological and social theory.

To this reviewer, the volume does put forth a good case for the use of psychological and social theories in its treatment of the process of

education in the undergraduate college. However, one is left somewhat adrift between the support of a theory or of a hypothesis and the application thereof. In few places is a theory stated explicitly, whereas hypotheses are fairly common. However, there is a great deal of penetrating analysis especially in Part VI dealing with human behavior and with its modification by means of human interaction.

The volume's thesis of the need for a scientific basis for action in higher education has a message for the educational researcher. The major shortcoming of educational research has been and still tends to be that it is largely "local and practical in its orientation." Thus, it should be scientific and not so locally oriented as it is. It should not be so highly directed toward school operations as it is. Thus, educational research must be concerned with human behavior and familiar with the areas of knowledge in the behavioral sciences.

Because of *The American College's* great emphasis on the individual, student, and teacher, little attention has been paid to the college as an organization. Professors Riesman and Jencks in Chapter 3 use case studies for an ethnographic or anthropological study of the college as a system. One wishes that a chapter had been written on the organization that is a college from the point of view of the administrative theorist.

Another major purpose of the book is to encourage behavioral scientists to make higher education a field of study. The editor speaks to this point as follows.

One might be tempted to speak of a "science of higher education," in order to accent the notion that the field may ultimately be constituted as a body of fact and theory, a discipline of sorts, in which individuals might become specialists. (p. 29)

This reviewer came away from the book with the feeling that Professor Sanford and his colleagues did indeed provide stimulation, through provocative and sometimes provoking discussion of the college, for the social scientist to look closely at higher education as a possible fertile field for research.

The references at the end of each chapter give some indication of research which has been done. Each author includes such a list of references with the exception of Anthony Ostroff (Chapter 12—Economic Pressures and the Professor), Harold Taylor (Chapter 23—Freedom and Authority on the Campus), and Frank Pinner (Chapter 27—The Crisis of the State Universities: Analysis and Remedies). Two chapters have a voluminous reference list: Chapter 8—Procedures and Techniques of Teaching: A Survey of Experimental Studies by W. J. McKeachie and Chapter 19—Dropouts from College by John Summerskill. In fact, a reader might reason, perhaps superficially, that the nine pages and seven pages of references, re-

spectively, in Chapters 8 and 19 indicate that the bulk of research in higher education has been on pedagogical techniques and on the student who left before completing his degree.

Higher education as a field of inquiry is certainly undeveloped, and *The American College* has highlighted this fact. Certainly more systematic scientific analysis needs to be done on the application of the present state of learning theory to higher education. Much more must be done in applying research in administrative theory to the organization of the college. As Professor Sanford says in his introduction:

In some areas it has been possible for an author of this volume to report and to summarize a large body of research. . . . In other areas the author has had to rely upon observation or clinical impression . . . (p. 3)

The lack of the former and the surfeit of the latter in *The American College* supports the volume's goal of emphasizing the need for a "... generally accepted theory of individual human development in accordance with which colleges may state hypotheses pertaining to the relations of ends and means." Two years have passed since the volume's publication, but the state of affairs today is much the same.

C. RUSSELL DE BURLO, JR.
Educational Testing Service (Princeton)

Conflict, Decision and Dissonance by Leon Festinger, with the collaboration of Vernon Allen, Marcia Braden, Lance Kirkpatrick Canon, Jon R. Davidson, Sara B. Kiesler, and Elaine Walster. Stanford, California: Stanford University Press, 1964. Pp. x + 163. \$4.75.

This small volume deals with three aspects of the decision-making process: behavior in the pre-decision situation, behavior following the decision, and the relationship between the two. The analytic emphasis in each case is on cognitive processes, and, as one might expect, dissonance theory provides the conceptual framework.

Following an introductory chapter, each of five chapters is devoted to a specific problem area within this realm: "The Difference between Conflict and Dissonance," "The Onset and Rapidity of Post-Decision Processes," "Seeking Information before and after Decisions," "The Post-Decision Process," and "Aspects of the Pre-Decision Cognitive Process." In each chapter, some speculations and hypotheses are presented, followed by reports of two experimental studies designed to test them. These experiments were performed by Festinger and several of his students (the collaborators) over the past three years, and apparently have not been reported else-

where. A final chapter sums up the major findings and discusses some of the remaining problem areas.

On the positive side, the book is concisely written and fast-moving; refreshingly lacking are the usual drawn-out examples and elaborations. The material as a whole is organized coherently and articulates nicely—a situation making for pleasant reading. Some of the ideas are intuitively compelling and may prove to be valuable contributions to the conceptualization of the decision-making process. Also, several of the experimental designs employed are quite ingenious and interesting.

The principal fault in this volume appears to be in the use of the experimental data. As is too frequently the case in experimentation, more emphasis is placed on the statistical significance of results than on their meaningfulness and substantive importance. Several conclusions are based on changes and/or differences in subjects' ratings which, although significant, are relatively small in relation to the amount of possible change, e.g., a difference in change of .30 units on a seven-point scale. Also, such changes or differences are smaller than one would expect if the processes which are hypothesized as producing them are as important as the authors believe. The frequency of such results seems to indicate, at the least, that uncontrolled factors are operative in the experiments, or that there are masking factors in the processes themselves which remain to be specified.

A second problem is the use of minor data to confirm hypotheses and generalizations when principal data fail to do so. The seriousness of such uses of results is increased by employing them, in some places, as apparently unequivocal support for major points or "critical" tests of hypotheses. (An exception is the experiment by Canon, in which striking differences between conditions were found.) Although both uses of data may be justified in specific cases, the authors should acknowledge that conclusions based on such results are less than fully confirmed.

On the whole, the book is worthwhile reading for those interested in the psychological processes of decision making. It provides a good, brief overview of some of the basic concepts of dissonance theory, the methodology used in testing its hypotheses, and the results typically obtained, for those who are unfamiliar with it. In addition, it contains some of the most recent experimentation and conceptualizations by Festinger and his colleagues, including revisions and extensions of the theory of cognitive dissonance.

JOHN DELAMATER
University of Michigan

Science of Educational Research by George J. Mouly. New York: American Book Company, 1963. Pp. viii + 515.

Educational research is not a hobby. The bulk of educational research is now being done by excellently trained specialists, instead of by administrators and classroom teachers seeking "practical" answers to immediate problems. The text that teaches that educational research is akin to casual data gathering no longer has a market. The text that has a market is one which exhorts aspiring researchers to train themselves in the scientific method and to begin the work of building a scientific superstructure on the base of existing unrelated empirical results and speculations. Mouly's text emphasizes that educational research can be a science.

The thesis of the text is that "... probably no obstacle stands so clearly in the path of the progress of the science of educational research ... as our failure to integrate the multitude of empirical findings which the reams of research studies have produced into meaningful structure." In the first of three parts of the text, "Science and the Scientific Method," Mouly instructs the reader on what a meaningful structure is and on how it is built and made more meaningful. Part II, "Research Techniques," contains the usual modicum of material on statistical methods. Part III, "Research Methods," is organized around "types" of research, viz., historical, survey, experimental.

Texts of this sort attempt to cover such diverse material (indeed, Mouly himself wonders whether educational research is a sufficiently homogeneous field that common training of researchers is possible) that each possesses numerous strengths and weaknesses. Travers' 1964 revision of *An Introduction to Educational Research* is perhaps the most even modern treatment of educational research. Mouly's text is quite spotty. It is strong in the area of scientific methods and in its evaluation of the present state of educational research; it is weak in statistics and measurement.

Students and professors alike will be impressed with the first 100 pages and with Chapter 14. In these pages Mouly develops the thesis of the book which was quoted above. His discussions of methods of reasoning, Mill's canons, causation, proof, hypotheses, and theories are quite scholarly. Students may find parts of this material redundant. Mouly generally fails to use illustrations from education in the first three chapters. In addition his "Projects and Questions" which follow each chapter are lifeless and probably will not appeal to education students; for example, on page 43 the student is directed to report on the development of the periodic table in chemistry. These shortcomings will not detract from the unique value of this part of the text, however.

Chapter 5 is a routine description of the library. Chapters 6 and 7 are devoted to statistics and sampling. The sampling chapter is better than most such chapters in similar books. These chapters are the weakest portion of the text. The discussion of statistics is out-

dated; the last 50 years of inferential statistics receive attention in one paragraph on page 156. The most egregious errors are the incorrect definitions of Type I and Type II errors on page 152: Mouly inadvertently reversed the definitions. Curiously, pages 179-180 comprise 5,000 random digits, though the readers are not told how to use them nor how they arose. Six pages on modern data-processing equipment partially redeem Chapter 6. To the reviewers' knowledge, Travers is the only other author who has given this important topic much attention. Mouly did not succeed in directing the students to those materials in statistics that would be useful to them. The entries under Selected References are very advanced works with few exceptions.

Measurement is given short shrift. Reliability is disposed of in less than 200 words; validity fares only slightly better. Any discussion of scales of measurement is missing.

Chapter 15 is exceptional. Mouly has abstracted 21 of the most important (for education) research investigations of this century. The methodology and findings of each study are summarized in two or three pages. All this makes for 50 pages of interesting reading.

The graduate research methods course in schools of education is growing "like topsy." Educational researchers are constantly employing new methods or adopting the methodologies of other disciplines. What is to become of the textbook that attempts coverage of all research methods? What will happen to the course that is supposed to cover "methods of educational research"? The reviewers feel that one course in research methods cannot prepare students to perform worthwhile research. The research methods course can no longer replace a sequence of measurement, statistics, and methods courses. Traditional methods courses should perform the function of instructing beginning graduate students in the principles common to the many facets of educational research, viz., the principles of science. Mouly's text does more instructing in the principles of science than any of the other texts generally used for the same course. In the reviewers' estimation, this feature recommends the text most highly.

However, those seeking a text for the traditional methods course which must serve as a substitute for two years of training will probably find *Science of Educational Research* too weak in measurement, statistics, and experimental design. Hopefully, the patent weaknesses of the text will not cause its strong points to be slighted.

GENE V. GLASS

Laboratory of Experimental Design
University of Wisconsin

DALE E. MATTSON

American Association of Dental Schools

Growing Up in River City by Robert J. Havighurst, Paul Hoover Bowman, Gordon P. Liddle, Charles V. Matthews, and James V. Pierce. New York: John Wiley and Sons, Inc., 1963. Pp. 189.

Growing Up in River City is a compact readable report of a longitudinal study which investigates the chief formative influences working on a group of boys and girls living in a medium-sized mid-western city. The authors' major purpose is to show the effect of forces such as social background, education, and personal characteristics on a group of children's achievement of tasks leading to competency or incompetency as young adults.

The investigators were concerned with the question: to what extent and under what circumstances do children who show promise of superior performance actually develop their talents?

The authors present their analyses of extensive observations of a group of 487 boys and girls who were used as a control group in a 1951-1960 University of Chicago research study designed to help "River City" improve its growing-up conditions for children. This control group was composed of approximately equal numbers of boys and girls most of whom were 10 or 11 year old sixth graders when the study began and 19 or 20 year old youths at the end of the study.

Case studies, clear concise tables, and a highly relevant 15 page appendix constitute almost half of the 189 page text which has a three-page index. As is typical of Mr. Havighurst, the senior editor's work, prediction of probable development, is an important part of the value of the book. Care is taken to support with statistical data statements pertaining to statistical techniques and tests used, and the significance level of the findings appear in the tables, text, or appendix.

Generalizations beyond the sample are included and not infrequently the reader may find some of these distracting. Probably the brevity of the book obviates documentation of statements such as "marriage is the only constructive behavior of which this type of girl is capable" or "marriage is the best solution" as appears on page 129.

As the writers discuss their "River City" findings relative to school drop-out, early marriage, delinquency, who goes to college, successful and unsuccessful work experience, they treat these timely topics as end results of formative influences on children in the study. They point out how combinations of forces lead to behavior that could be predicted.

The book has an optimistic point of view and contains suggestions on how communities can organize to give youth a good chance to grow up into competent adults.

Some of the most significant findings and conclusions drawn from the study include the following.

1. The best equipment for satisfactory growth is to have a keen mind, to accept oneself and be well accepted by others, and to come from a middle-class family.

2. A large minority of children from lower-status families do very well in school and community; many rise above the social class level of their parents.

3. Boys and girls who do well in school are those who in addition to their superior mental ability generally have families that help and stimulate them to do good school work.

4. Patterns of delinquency are often evident at age 10 or 11. A first sign of trouble in growing-up is school failure demonstrated in poor reading and some form of aggressive maladjustment.

5. There is a large group of boys and girls who are unsuccessful in becoming competent young people with whom the church has no contact.

6. The study shows a strong tendency for those who are better adjusted personally to attend college. Although three of the 92 most severely maladjusted children did attend college, these three came from middle-class unbroken families, the top quarter in I.Q. and school marks. These factors seem to offset the maladjustment.

7. Motivation for college attendance differs in girls and boys. While college attendance is an on-going continuous growing-up plan for boys, educational motivation of girls shifts at the end of high school as marriage and motherhood take precedence. This change causes conflict in girls of high intellectual ability and achievement drive. There is little relationship between college attendance and social class of girls in this study. Values and personal choice play a greater part in girls' college attendance than in boys'. Of girls in the upper quarter of intellectual ability, those well motivated for college had less drive for achievement than those not motivated for college.

8. Prediction of work adjustment can be made with substantial statistical reliability from I.Q. at sixth grade level and socio-economic status.

9. Adult competency can be predicted from combining data on school achievement, personal-social adjustment, and family social class.

10. Six distinct adjustment groups each with its characteristic pattern of growing-up emerged from the observations.

11. A good school system builds toward a cohesive society by preparing boys and girls to find places satisfactory to them in adult society.

12. "River City" needs a system of differential opportunities such as work training as an alternative to school. This is particularly needed by disadvantaged boys and girls as a pathway to adulthood.

ALMA V. WILLIAMS
Laguna Blanca School

Television and Human Behavior by Leon Arons and Mark A. May (Editors). New York: Appleton-Century-Crofts, 1963. Pp. 307. \$7.50.

The study of the problems of television as they are specifically related to the impact of this medium upon the American audience grew out of a challenge to social scientists from education, industry, and institutions of public welfare to create specific research plans underlying the solution of a large number of problems related to television and its impact upon human behavior.

Competition for such plans, supported by the television industry, was announced in the summer of 1960. Social scientists who were expected to submit plans were encouraged to view television in the broader cultural environment of modern America. The specific purpose was to produce plans of research either for television itself or in relation to other means of communication. This announcement specified that the concern of the governing committee was with television and human behavior whether psychological or social. The plans submitted were to pertain either to basic or applied problems involving experimental studies, surveys, correlational studies, methodological developments, conceptional matters, and communication theory, or any combination of these.

It was hoped that the competition would inspire participation of competent persons in universities, business, industry, government services, and other fields. This hope was realized. One hundred forty-seven plans were submitted of which 130 met the committee's general specifications. In the announcement of the competition, awards were promised for 18 plans to be selected for publication in a special book. The authors were to receive an honorarium of \$250 for each plan selected. The authors of the two outstanding plans would be awarded \$4,000 for the first place and \$1,500 for second place.

Television and Human Behavior, published by Appleton-Century-Crofts, presents all of these plans with other details of the project. The scope of these problems was indeed impressive, ranging from such general problems as mass media and interpersonal communication, the immediate and remote effects of television, and a wide variety of concerns for the consequences of program telecasting to more specific concerns involving persuasion and the initiation of action through policy intervention as well as embracement of such current problems as income tax compliance, the logic of politics, the consequence of television advertising, the effect of television in inducing action, and the instrumentation for behavioral research. It was interesting to note that television, as related to education, was involved in three of the research plans submitted.

This book should hold special interest for those concerned with progress toward systematic performance of intellectual tasks in-

volved in solving social problems. Many are concerned also with the creation and appraisal of alternative interventions in current television programming which many feel contribute to adolescent delinquency. It holds special interest for educators who realize that a program of teaching the methods of systematic thought has been widely neglected not only in secondary schools, but also in institutions of higher learning.

Plan Number 18 by Drs. K. V. and W. M. Smith received the second award. This plan made a distinguished contribution in bringing together the available scientific television methods of analyzing perceptual feedback in behavior.

The appraisal committee selected for first award Plan Number 7 by Arthur J. Brodbeck and Dorothy B. Jones. This plan deals with the purpose of this entire project which was effective advancement in the quality of research in the solution of problems of television in its interactions with human personality. For a number of reasons this award was preeminently justified.

The plan was based upon a comprehensive intellectual framework which can be employed in a consistently systematic manner. This approach proposed five intellectual tasks as fundamental to the solution of the specific problem of determining the effects of television viewing upon the norm-violation practices and perspectives of adolescents. These intellectual operations were clearly stated as goal clarification (statement and justification), trend analysis (of past events as they pertain to the goal), condition analysis (of present events as they tend to hinder or contribute to the realization of the goal), projective estimates of probable events of the future as they relate to the goal if existing conditions persist, and the creation of alternative interventions designed to supplant tendencies which hinder progress toward the goal with practices designed to facilitate the goal.

It is to be recognized that the justification of any goal must be based upon a predisposing orientation of the researcher. Unfortunately, this orientation often represents an unconscious predisposition and leads to the intuitive selection of the goal which therefore is not subject to specific justification. In the case of the Brodbeck-Jones Plan, however, the conscious orientation was specifically toward the wide accessibility of human values necessary to the realization of human dignity. Furthermore, these values can be defined in terms of specific reference, and with the clear understanding of the degree to which institutional practices may restrict or facilitate their wide accessibility. The clarified problem then becomes one of systematically gathering intelligence about the influence of television upon the value statuses of the young and the development of appropriate strategies designed to facilitate their enhancement on a grand scale through this mass medium of communication.

The specification of the intellectual tasks of problem solving of-

fers the student of the future a thrilling prospect. The question of the creation of alternative strategies designed to achieve the goal lends itself to value analysis in a number of ways. Delinquent behavior can be classified in terms of the value specialization. Appraisal effort can be organized in terms of value outcomes. Value categories proposed in the plan provide a widely applicable intellectual tool for the analyses of the needs and wants of human beings in all cultural environments, of the frustrations of everyday living, of the degree of the realization of potentials of individuals and groups of their perspectives (demands, expectations, and identifications), of the outcomes of their behavior, and of the impact of intervening strategies used in efforts to modify behavior.

The authors make very clear how value theory can be employed in a synthesizing role. They have performed brilliantly in proposing a technique of systematic inquiry that has many applications in the social sciences which have not been widely employed except in the field of political science by Harold D. Lasswell and his associates.

This approach can be found to be especially useful in providing many intellectual instruments which can be employed in attacking the problems of education. For many years, fundamental problems of education have gone unanswered largely because of the failure to employ a science of values and appropriate frameworks of inquiry such as those proposed by the authors of the award-winning plan. These instruments will facilitate the pursuit of such problems through techniques that preserve intellectual orientation at every step in the performance of the intellectual tasks necessary to their solution and to decision making upon which the creation of enlightened educational policy depends.

The value-oriented approach provides explicit perspectives for an educational philosophy compatible with democratic ideals. It embraces the following concepts. The overriding objective is the realization of human dignity. Important goals are the wider sharing of human values. The formation of mature personalities whose capacities and demands are compatible with these ideals must be facilitated. The long range goal of the educative process is to provide opportunities for the individual to achieve his full potential. The student must be led to seek these values for himself with minimum damage to the freedom of choice and the value assets of others.

The synthesizing role of value theory suggests a highly effective approach in performing the task of integrating the various disciplines of the educational program which have become so highly splintered through increasing specialization. The fact that value categories can be used to classify the impacts of all the disciplines upon the distribution of human needs and wants (values) has been largely ignored as an integrating factor so long sought in the construction of educational programs.

Lip service has been given to the goal of teaching students to

think, but little has been accomplished in actually undertaking any steps to achieve this goal. The authors of this plan deserve commendation for their creation of a plan for transmitting to students the skills of systematic thinking toward valid goals.

It would seem to be imperative that those responsible for the creation of educational and psychological tests and measurements forthwith assume responsibility for exploring the intellectual approach presented in the Brodbeck-Jones Plan in its application to the validation of educational goals and in the appraisal of educational practices. If this is done, the future will reveal that the award-winning plan presented in *Television and Human Behavior* has created a distinguished landmark in the history of education.

V. C. ARNSPIGER

East Texas State College

The Quick Test (QT): Provisional Manual by R. B. Ammons and C. H. Ammons. *Psychological Reports: Monograph Supplement*. I-VII, 1962. Pp. 51. \$2.50.

This excellent monograph by Ammons and Ammons is purported by the authors to be a provisional manual for the *Quick Test (QT)*, a quick screening instrument for measuring verbal-perceptual intelligence; but it is much more than that. It is a combination of (a) an exposition on the merits of brief screening devices for estimating a wide range of human intellectual abilities; (b) a plea to critics of short psychological tests to consider factors other than brevity in their evaluation of such instruments; and (c) a review and summary of the professional literature reporting a great scope of research findings with the *QT* and its parent test—*The Full-Range Picture Vocabulary Test (FRPV)*. There is also a regular test manual to accompany the *QT* as well as a suggested list of further research to be conducted by the user of the *QT*.

The *QT* is published in three single forms comprised of 50 word items, each of which can be administered in two minutes. All forms can be administered in six to ten minutes. A short version of the *FRPV*, which itself is a short test as measures of intelligence go, the *QT* can be administered in 10 to 12 minutes.

The authors begin the discourse contained in this rather lengthy monograph by asking the question: why a new test? They answer it by stating that they feel that there are significant advantages which accrue to having new stimulus drawings, new independent forms of the *FRPV*, and very short tests made available to users for whom time is an important factor. They then go on to state their purpose for preparing this provisional manual; namely, to provide users with a standard set of instructions and interpretative materials for the *QT*—and also (which perhaps reflects their basic motivation in preparing such an unusually comprehensive manual) to provide a

rebuttal for critics of the *FRPV* who have maintained that it is inadequately standardized as well as insufficiently and unconvincingly researched. On this point the authors state in the monograph (with some documentation) that there has actually been so much research conducted with the *FRPV* that they (the authors) have not been able to keep up with it all. Consequently, in their attempt to keep abreast of the research findings with their instrument, they neglected to prepare a comprehensive manual for the *FRPV*. Hence, in order to avoid the occurrence of the same situation with regard to the *QT*, they have published this comprehensive provisional manual as a monograph supplement *prior* to releasing the *QT* for operational use. For this step, the authors are to be commended, and perhaps in this action they may be to some degree absolved of any previous "sin" they may have committed in the eyes of their severe critics in the publication and distribution of the parent test, *FRPV*, without a proper manual. However, to their defense on this point, the authors refer to the numerous publications of articles in the professional journals reporting the developmental and operational research results on the *FRPV*.

The authors fully and adequately describe the developmental research and standardization procedures for the *QT*. Separate norms for children, young adults, and older adults are presented and discussed. The authors advocate and plead for the construction and use by the consumer of locally prepared norms. In reporting the reliabilities, they summarize 10 separate studies; the sample *n*'s range from 20 to 100 cases. The reliability coefficients based on comparable forms range from .60 to .96. They are reported as reliability estimates determined from the mean *z* between and among the various forms.

It is refreshing and reassuring to this reviewer to observe that a proper procedure (comparable forms) was used in every study in establishing the reliability estimates for the various test forms. Too frequently, authors are prone to use the inappropriate internal consistency procedures which only lead to highly inflated reliability estimates for speeded tests. It is further gratifying to realize that reliabilities as consistently high as those reported in these 10 independent studies are likely to result only if the scores on these brief tests are based on a very few, but highly discriminating items.

One further brief word about the reliabilities: predicted reliability estimates are given for 2-form and 3-form tests, determined on the basis of applying the Spearman-Brown formula. These may be spuriously high, but then so are likely to be all correlations corrected by this formula. At least the basic coefficients are the correct ones.

The validities reported for the *QT* are quite numerous; there are 74 validity coefficients reported in a summary of direct studies of

QT validity. However, they are for the most part concurrent validities, although this situation is not unusual for tests purporting to measure human intellectual abilities. The principal criteria in these studies yielding validity coefficients ranging from $r = .13$ to $r = .96$ have been such tests as the *Full-Range Picture Vocabulary Test* itself, the *Ohio State Psychological Examination*, and the *Iowa Test of Basic Skills*. Such other criteria as final school grades in social studies, reading, spelling, English, and arithmetic subjects have been used in a few studies. However, in these studies the reader is not given the information upon which to determine whether or not these represent predictive or again merely concurrent validity coefficients. Perhaps it would be worthwhile for the authors to clarify this situation for the reader in their subsequent regular manual. The validity coefficients are based on relatively small samples, ranging in size from 16 to 80 cases. The same situation also exists for the reliability and normative studies reported here. A quick judgment would be that the small sizes of the samples employed in these studies are likely to cast doubt upon the meaningfulness of the results obtained. However, as the consumer explores more deeply into the nature of the samples and recognizes the wide range of subjects used: Negro, American Indian, Spanish-American, disturbed children, aphasics, and others; and as he comes to realize the wide variety of controls effected with these samples such as those for age, sex, grade placement, occupational level, geographic location, he will likely acquire considerable respect for the outcomes of these studies.

Considerable attention is focused in the monograph upon the training and qualifications of the user of the *QT*. Although the authors state that there is no substitute for training and experience, and that ideally no independent testing of this nature should be done by anyone who does not have a doctorate in clinical psychology and several years of supervised experience, they discussed several practical considerations. From a practical standpoint, they maintain that they have found that reasonably intelligent adults with no formal training in testing can learn to administer the *QT* efficiently and that with some additional training such persons can be taught to interpret it adequately. They go on to say that it is better to train non-psychologists to administer and interpret these tests adequately rather than unrealistically to expect untrained persons to do absolutely no testing! In line with this reasoning, they have simplified the presentation of directions and materials in this manual. Such statements will undoubtedly elicit some strong negative reactions on the part of the authors' professional colleagues, but this reviewer for one, on the basis of his own personal experiences, is willing to agree with Ammons and Ammons.

In summary, because many of the approaches indicated and the statements made in this monograph are likely to be perceived as un-

conventional and, therefore, erroneous by many readers, the controversial nature of the contents of much of this provisional manual for the *QT* is deemed inevitable. There will be those who will undoubtedly be greatly concerned about the brevity of the tests, the relatively small samples utilized in the normative, reliability, and validity research, and the rather high correlation coefficients reported which suggest spuriousness. To these critics, and to all users or potential users of the *QT*, this reviewer can only advise others to consult the basic research literature on the *QT* to which the authors make repeated reference in the monograph. Of course, this reviewer is not necessarily willing to accept all the findings reported in the monograph at face value. Nevertheless, the data and arguments presented by Ammons and Ammons are both impressive and seemingly convincing. Therefore, they cannot be blindly ignored!

If the *QT* is only partly as good as the data and findings reported in this provisional manual seem to imply, then the authors will have made an outstanding and lasting contribution to the field of psychological testing. It is necessary, however, for the discriminating user and researcher to go beyond the data reported herein and, as the authors themselves suggest, conduct his own research with the *QT*.

PETER F. MERENDA

University of Rhode Island

Theory and Research in Projective Techniques (Emphasizing the TAT) by Bernard I. Murstein. New York: John Wiley & Sons, 1963. Pp. xiii + 385.

In 1938, Murray and his coworkers published their exciting inquiry into the nature of man. They asked the question, "What are the fundamental variables in terms of which a personality may be comprehensively and adequately described?" They found thematic apperceptions a provocative technique for their purposes, and the *Thematic Apperception Test (TAT)* was born. As a vehicle which has promoted deeper understanding of man's nature, its historical position is unassailable. However, its value as a clinical instrument is less certain. The reliability of interpretations made by average competent clinicians has been found wanting, and the *TAT* is under attack along with other projective instruments.

Every reader knows of gifted clinicians who have the skill to delineate that "X-ray picture of the inner self" by selecting those cues in a protocol which are essential, and by rejecting those which under the given circumstances are not applicable; but their techniques are not so consistently available to the less gifted. However, lesser clinicians still find projective methods valuable because of the possible insights which they provide, and so are loathe to discard them. How to transform the idiosyncratic methods of the genius into public methods is an urgent problem. Furthermore, as Murstein points out,

"the *Zeitgeist* is heavily accented with quantification, standardization, and control of the variables influencing the thematic response," and he warns that "if none of the attempts to provide the *TAT* with an appealing analytic-quantitative superstructure succeed, it may find itself relegated to a viable but minor role in the clinical armamentarium between the *MMPI* and the *Szondi* tests."

This book may be regarded as one important step in the effort now being made to find "an appealing analytic-quantitative superstructure." Approximately half the book is devoted to three of its thirteen chapters which deal most directly with this problem: Chapter 5, "Theories of Projective Techniques"; Chapter 6, "Reliability"; and Chapter 7, "The Stimulus." The discussion of theories touches briefly on psychoanalysis, and then turns to those approaches which invite quantification, such as the approach-avoidance conceptualizations and McClelland's scoring of the achievement motive. Since Murstein's own preference is for the field approach, he emphasizes the importance of interaction among stimulus, background, and organismic variables.

Accordingly, when he turns to examine statistical methods of testing reliability he places high value on analysis-of-variance procedures for the *TAT*. He also stresses their advantages over non-parametric procedures. Studies employing the various methods of testing reliability are didactically presented and critically reviewed. Thus, a reader who has general sophistication in this area, but who lacks immediate acquaintance with the methods being discussed, has no difficulty in following their rationale.

The chapter on "The Stimulus" takes on special importance because Murstein's own major contributions to the experimental literature on the *TAT* have dealt with the stimulus, which he finds to be, statistically, the major determinant of the content of the response. This chapter covers a wide range of topics, such as animals vs. humans, the confounding of structure with ambiguity, problems in the measurement of ambiguity, the role of socioeconomic factors as depicted on the cards, the need to have the entire range of a stimulus represented, and methods of improving the usefulness of the *TAT*.

Anxiety for the future pervades the book. The clinician immediately senses it in the expansive title which would seem to promise a treatment of all the major techniques, and which then is withdrawn into the protective arms of the apologetic parenthesis "(emphasizing the *TAT*)." The book in fact deals almost exclusively with thematic apperceptions, other projective methods being presented only as background. In a larger sense, the anxiety is justified and can be and should be quite salutary.

This book is essential reading for all concerned with the use of the *TAT* as a clinical or research tool. It is a valuable reference book for

the researcher; it can be used as a textbook for students in clinical programs—not however for learning how to administer or interpret the test—and can be employed as an aid to the clinician in deepening his awareness of limitations and possibilities of the TAT.

FLORENCE DIAMOND
Pasadena, California

The Morgan State College Program—An Adventure in Higher Education. by Martin D. Jenkins (assisted by seven collaborators). Baltimore, Maryland: Morgan State College Press, 1964. Pp. v + 104. Paper-bound, \$2.00.

Aided by seven members of his staff, Jenkins, a psychologist (Ph.D., Northwestern University, 1935) who has been President of the predominantly Negro liberal-arts Morgan State College at Baltimore since 1948, reports eloquently and candidly the many important things that he and his 170 faculty members are doing for their 2500 regular-session undergraduates. With changing conditions in our society, the predominantly Negro colleges are faced with the problem of meeting the needs of students who have experienced cultural and educational deprivation. Morgan State College has developed a multi-faceted program that has produced remarkable results.

On pages 6-9, President Jenkins lists 68 "elements of the educational program" at Morgan State College, many of which are especially novel and ingenious. Most are remedial or enriching in order to compensate for the cultural deprivation which the majority of these students have experienced all their lives. Some of the most carefully worked out are a three-track freshman program, a revised freshman English program, a required course in reading for all freshmen, an elective course in vocabulary building, extensive testing and counseling, and seminars on how to take tests.

Though Jenkins's primary topic is curriculum, frank concern and puzzlement about standardized testing permeate the book. He is no stranger to psychometrics, as his "The Upper Limit of Ability among American Negroes" (*Scientific Monthly*, 1948 LXVI, 399-401) and "Differential Characteristics of Superior and Unselected Negro College Students" (*Journal of Social Psychology*, 1948 XXVII, 187-202) reveal. It is tempting, however, to believe, because on the average Negroes tend to score lower on scholastic aptitude and achievement tests than do non-Negroes in their geographical locality, that such tests must therefore be "invalid": "... it is well known that standardized examinations have low validity for individuals and groups of restricted experiential background" (p. 93). This does not seem true within predominantly non-Negro colleges, however, so far as the limited amount of evidence known to the reviewer indicates about the prediction of freshman grades.

The most complete publication known to the reviewer on this issue¹ shows the median multiple r for predicting average freshman-year grades from the best-weighted linear composite of Scholastic Aptitude Test (SAT) Verbal and SAT Quantitative scores as .52 for men and .57 for women at three predominantly Negro Georgia state colleges (Albany, Fort Valley, and Savannah), versus .41 and .57 respectively, at the predominantly non-Negro Georgia state colleges (15 for men, 14 for women).

Because SAT-V and SAT-Q scores are much less variable in the predominantly Negro Georgia state colleges (these tests are too difficult for most of their students) than in the predominantly non-Negro ones, the standard errors of estimate in the former (median .52 for men and .48 for women) are lower than in the latter (.61 and .60, respectively).²

An official of the American College Testing Program (ACT) stated in the reviewer's presence at a conference in August 1964 that he had examined ACT data for similar colleges and found that standardized tests predict average freshman grades better in ACT's predominantly Negro colleges than in ACT's predominantly non-Negro colleges. Of course, his data and Hills's alone cannot *prove* the point, because data for other colleges or other years *might* tend in the other direction, nor are these data necessarily relevant to Negroes who attend predominantly non-Negro colleges, but they lend no support to the statement by President Jenkins.

Because Morgan State College students tend to fear standardized tests (e.g., "the required examinations [for governmental positions] tend to serve somewhat as a deterrent [to applying for them]," p. 78), much effort is made to help them acquire test know-how and test wiseness. Much of this is frank coaching on presumably similar materials. For example, "one hundred copies of *How to Pass High on the Graduate Record Examination*, by Edward C. Gruber, [were] placed in the library for use by students. Fifty copies [were] distributed to faculty members. [p. 8] . . . Students are provided with a variety of materials including reading guides, test questions, problems, and typed or recorded lectures. Chief emphasis is placed on the content of the major [i.e., the advanced subject-matter] examinations of the Graduate Record Examinations" (p. 52).

Undoubtedly, such efforts are needed and should not be discouraged. Their results should be evaluated carefully, preferably by

¹ Hills, John R., Klock, Joseph A., and Lewis, Sandra C. *Freshman Norms for the University System of Georgia, 1961-62*. 244 Washington St., S. W., Atlanta 3, Ga.: Office of Testing and Guidance, Regents of the University System of Georgia, 1963. Pages 4 and 1.

² Medians for the academic year 1962-63, which became available after this review was prepared, are as follows: multiple r 's, .52 and .45 versus .42 and .54; standard errors of estimate, .52 and .50 versus .59 and .54.

controlled experiments, however, because most studies of the effectiveness of coaching—particularly at Educational Testing Service—give little basis for optimism. Among culturally disadvantaged students the results *might* be different, but Morgan State College needs to know what kinds of coaching, and how much, are needed, so that much possibly time-wasting activity can be avoided.

The evaluation staff at Morgan State College has tried valiantly to estimate the effects of coaching, but so many variables operate differentially from year to year there that it is all too easy for a reader of this volume to propose alternative explanations of increases in percentile ranks on standardized tests. One will want especially to study pages 47-49, "Student and Institutional Evaluation through Tests," and 51-53, "Efforts to Up-Grade Student Test Performance." President Jenkins is by no means naive about these matters, as his frequent comments throughout the book indicate.

On pages 78-82 Jenkins mentions a few aspects of the background of the 134 graduates of the College who were awarded graduate fellowships or assistantships at 57 universities during the period 1957-63, when "21 graduates are known to have attained academic doctorates, in addition to 20 who earned doctoral degrees in medical fields" (p. 81). Table 4 shows that these 134 as entering freshmen had a median score at the 17th percentile of national college-freshman norms on the *American Council on Education Psychological Examination* or the *School and College Ability Tests*, with 83 of them (i.e., 62 percent) below the 25th percentile. No information specifically about the 21 academic doctoral recipients (6 of them in 1962) is given, nor are *GRE* scores for any of the 134 presented. It would be helpful to know them, as well as the academic records of these 134 students at Morgan State College and in their graduate work. Some ratings by graduate schools are analyzed, however, and from them it would seem that many of the graduate students are doing well at excellent universities.

This frank report by the unusually able, hard-working president of a college serving its predominantly academically-disadvantaged students exceptionally well deserves the careful attention of most measurement specialists, because testing these students and evaluating Morgan State College's efforts call for measurement skills of a high order, and it is essential reading for all persons interested in the education of such students. President Jenkins's "A Word of Interpretation" eloquently indicates how generally applicable his program is.

The point of view expressed here calls for the addition of a compensatory and enriching overlay as remediation for constricted cultural participation. By no means does it imply the necessity for a special kind of education based on race. The phenomenon cuts across race lines. The principles enunciated, con-

sequently, are applicable to any students who enter college with serious educational and cultural handicaps.

Nor does this point of view call for or result in a sub-standard educational program. Basic academic remediation, largely restricted to the freshman year, demands additional time of students without infringing on standard college requirements. The functional emphasis on general education—as contrasted with mere academic knowledge—is universally applicable to institutions of higher education.

Viewed in this light, the Morgan State program has implications for other than predominantly Negro institutions of higher education.

JULIAN C. STANLEY

*Laboratory of Experimental Design
University of Wisconsin, Madison*

Guidance Services in the Modern School (Second Edition) by Merle M. Ohlsen. New York: Harcourt, Brace & World, 1964. Pp. 515.

The general orientation of this book is practical. It is designed to help teachers, administrators, counselors, and prospective counselors to understand the role of guidance services and counseling techniques based around a child-centered philosophy relevant to pupil placement, diagnosis, assessment, prediction, and evaluation.

This new edition devotes considerable space to basic principles of psychological measurement. The attempt to improve educational examining is made available through a presentation of fundamental statistical concepts and techniques necessary for summarizing and interpreting test results or scores as well as for showing their logical bases and their interrelationships.

The treatment of measurement techniques is confined to minimum essentials, involving application of statistical concepts relating to percentiles, correlation coefficients, *Z*-scores, *T*-scores, and stanines. No prerequisite knowledge of statistics is assumed. Perhaps its principal purpose is to expedite the transition from theory to practice of psychological testing techniques in terms which teachers and counselors can find most helpful.

This edition reveals the wealth of experience and sound scholarship of the author. It incorporates up-to-date research and reflects nine additional years of experience in counseling and guidance services, teaching guidance courses, and undertaking research.

Among the new topics dealt with are the ethics of testing, the underachiever, the exceptional child, theories of vocational choice, problems girls face in achieving vocational maturity, the implications of theory and research for vocational counseling, and the nature of social and leadership development in guidance programs.

This is a resource book that counselors and school personnel should find to be a worthwhile investment.

MABEL E. HAYES

University of Southern California

Guidance Techniques for Elementary Teachers by Ralph Garry.

Columbus, Ohio: Charles E. Merrill Books, Inc., 1963. Pp. 356.

\$7.50.

This book is definitely the work of a practitioner rather than a theorist. It is exceptionally readable and geared to daily classroom use. As the title implies, it is for the use of elementary teachers, but just as assuredly could be the springboard for the person desiring to enter the field of counseling. The treatment of the entire area of counseling, and of all the ramifications therein, is comprehensively undertaken without the introduction of superfluous detail.

The section on testing is especially attractive in that no violation of integrity of the trained teacher is involved. There is little chance for needless duplication and redundancy either of other course work or of previous pages in the book. This section contains a valuable series of tables in which the author lists the following information about tests: name of test, author, publisher, cost, time needed to administer, validity, reliability, and evaluation of the test. All tests included are applicable to the elementary level, and the list is not meant to be all-inclusive.

Another commendable feature is the almost complete lack of case studies. The book is devoted to the down-to-earth procedures involved in obtaining pertinent information, in evaluating and effectively communicating such information, and in using such information to chart an action course to better the environment for all concerned.

The book is so thorough as to have given attention to the exceptional child. So often texts deal with the norm at the teacher's level that it is gratifying to see the inclusion here. The disorders of the marginally handicapped will be a revelation to many teachers. This last section adequately and forcefully demonstrates how teachers can be valuable in early recognition and therapy.

This reviewer would have liked the inclusion of a short section on outside agencies that are often involved in aiding the schools. Many teachers are unaware that fraternal, business, community, and private organizations stand by to furnish glasses, hearing aids, clothes, and even jobs, to aid children with problems.

All in all, this is one of the best texts seen to date in terms of overall, usable, and professional data found on the market today. This book represents a real service to the profession.

ROY M. FITCH

San Fernando Valley State College

Interpersonal Dynamics: Essays and Readings on Human Interaction by Warren G. Bennis, Edgar H. Schein, David E. Berlew, and Fred I. Steele. Homewood, Illinois: The Dorsey Press, 1964. Pp. xv + 763. \$8.50.

This book is well-suited to fit three types of individuals concerned with interpersonal relations. These are the teacher, the student, and the intelligent layman.

The contributors do not draw their points of view solely from the field of psychology. They call upon the arts, literature, philosophy, and other disciplines to make their attitudes known in a refreshing manner. Symbolic reactionism, Sullivanian psychiatric theory, object-relations theory, and existential psychology have all had a hand in the development of this volume.

The essays are grouped into five major headings, which indicate the scope of this contribution. These are "Emotional Expressions in Interpersonal Relationships," "Some Interpersonal Aspects of Self-Confirmation," "Personal Change through Interpersonal Relationships," "The Instrumental Relationship," and "Towards Better Interpersonal Relationships."

Although it is difficult to do justice to this volume in a review of this length, a sampling of some of the pertinent remarks should point up its flavor. Kurt Wolff's essay concerning "Surrender" includes this statement: "Man, the being who can surrender and catch or invent, is inventive, as well as capable of being invented." The implications of this statement are far-reaching and go beyond a mere cognitive experience.

Harlow's discussion on "The Heterosexual Affectional System in Monkeys" offers many insights into the human affectional system, while Peter Lomas provides some rather interesting observations on "The Concept of Maternal Love," in which he discusses the relationship between narcissism and masochism. "The similarity between the two concepts is, however, more fundamental than the difference."

In the essay, "The Process of Understanding People," Tannenbaum, Weschler, and Massarik take up a discussion on Colin Wilson's *The Outsider* that "seeing too much" without having an appropriate range of behavior skills may mean the inability to function well within the realistic requirements of one's environment. Here, the authors point up the task of executives. Every executive faces responsibilities. He must learn to see clearly and accurately the human and inanimate factors of the total situation in which he is involved. In addition, he must acquire the skills to act in a manner which will tap wellsprings of positive behavior—a form of behavior, which, in turn, will eventually lead to a successful attainment of personal as well as organizational goals.

Carl Rogers in "This is Me" emphasizes the importance of only

being oneself. This statement means being able to accept oneself with imperfections. One of the interesting facets of this phenomenon is the fact that accepting oneself can lead to a meaningful acceptance of others. Rogers' writings point to the dynamic understanding of the flowing and changing process of life.

In the final essay, "The Teacher as a Model," Joseph Adelson discusses the teacher's role in the priestly mode and points out the problems raised. One of the items upon which he touches is that of "moral" questions. He does not imply "pleasure-seeking" connotations, but qualities such as integrity, fairness, courage, and ethical sensitivity. An interesting feature of this discussion is the consideration of the student in his relationship with the teacher.

Again, it must be kept in mind that only a few of the highlights have been cited. There are approximately 50 essays in this volume. Hence, there is enough for all interested students of the social and behavioral sciences. This book is appropriate for courses in any of these areas.

ARTHUR LERNER
Los Angeles City College

Determinants of Infant Behaviour II by B. M. Foss (Editor). Foreword by John Bowlby. New York: John Wiley and Sons, Inc., 1963. Pp. xii + 248.

This volume includes the proceedings of the Second Tavistock Seminar on Mother-Infant Interaction held under the auspices of the Ciba Foundation at the house of the Royal Society of Medicine at London in September, 1961. The Conference was unusual in that it considered the general subject of children and mothers from the aspects of several disciplines. Priority was given to empirical studies, especially those that utilized firsthand observations of what actually happens between infant and mother. *Determinants of Infant Behaviour I*, the proceedings of the first seminar also under the auspices of the Ciba Foundation, was concerned with the infant in his first year.

The ten reports in this volume are classified under the headings of animal studies, human studies, and method and theory. The two animal studies report on early experiences on mothering behavior in monkeys. The four detailed reports of mother and child relations include those of both American and African families. The four studies on method and theory are concerned with the techniques of studying infant behavior and with theories of developmental psychology. These studies are likely to be of interest to advanced students working in the disciplines of psychology, sociology, education, anthropology, and zoology.

The discussions that followed the reading of the reports at the

conference have been printed after the reports themselves. Many fruitful ideas can be found in these discussions.

The contributors—all leaders in their respective fields—represented psychologists, psychoanalysts, and zoologists. On the whole the studies were very well done, but vary somewhat as can be expected in any collection. It is extremely difficult to review simultaneously ten studies that are so widely diversified in objectives and methods as these even though they are on the same topic. Space does not permit any individual detailed review of each of these investigations. Therefore, additional comments will be limited to a few points.

A reading of both the animal and human studies provides the reader with some interesting comparisons between these two as well as suggests the possibility of a number of follow-up studies. The section on methods and theory points to numerous approaches to the study of young children. Many practices frowned on in child care were demonstrated to work in these individual studies.

JEROME E. LEAVITT

Portland State College (Oregon)

Educational Psychology (Second edition) by Karl Garrison, Albert Kingston, and Arthur McDonald. New York: Appleton-Century-Crofts, 1964. Pp. xvi + 554.

This textbook provides a refreshing change in orientation from many educational psychology texts. Although the field of educational psychology must of necessity be more applied than some other areas within psychology, it is the writers' opinion that the field is still basically an academic discipline. This text tends to present the field as such.

The book is divided into five parts and a total of 19 chapters. Part I provides a good introduction to the field as well as an introduction to the organism and his environment. Part II presents developmental aspects; in its extensiveness and depth it is in itself a brief course in child psychology. Part II is concerned with learning; the approach here is relatively eclectic. The authors present three major theories of learning (conditioning, trial and error, and insight) and then proceed to describe the classic elements in learning. Parts IV and V deal with the very practical aspects of evaluating pupils, studying children individually, coping with personalities and adjustment problems, and the like. Fifty pages are devoted to theoretical and practical aspects of measurement and evaluation. This material is cogent and up-to-date. If it may be assumed that the readers of this book retain the material to any extent, they should be more enlightened in the area of pupil evaluation than are the typical elementary and secondary school teachers today.

In addition to the usual list of selected readings, the authors have

included a number of problems and exercises at the end of each chapter. Somewhat unique is the listing of films correlated with the various subject-matter parts of the text. These films and their sources appear in the appendix.

The text is well organized; the titles and headings of various sections are meaningful and brief. The style of writing is clear and uniform in spite of the fact that the book has three authors. Illustrations are carefully chosen and well placed. The authors have also been able to achieve a good balance between presentation of studies and understanding of the subject-matter. Studies are not quickly passed over in the text, but rather receive an adequate description with the added inclusion of a table or graph when necessary.

In summary, this text appears to have achieved a proper marriage of the subject-matter and practical aspects in the field of educational psychology. It well organized and up-to-date. It is not "preachy," nor is it "sugary" (à la "understanding your pupils in the challenge of today"). It could be used not only by prospective teachers, but also by any group of students interested in the field itself.

PHILIP S. VERY

University of Rhode Island

Problems and Methods in High School Teaching by Adam M. Drayer. Boston: D. C. Heath and Company, 1963. Pp. xiv + 303.

The book is just what the title says, a problems book. Designed for use in courses preceding or accompanying student teaching, it emphasizes problems likely to be encountered in this important phase of the teacher's training. The short problems, numbering over 100, are pointed and real: they move to the core of a variety of problem situations likely to be encountered by student teachers—discipline, testing and grading, individual differences, motivation, and emotional adjustment, all treated in separate chapters. Additionally, the student is given paternal advice in such areas as personal hygiene and mannerisms, voice control, and social skills required for effective relationships with cooperating teachers, administrators, and pupils. He is also advised to be neat, courteous, punctual, to act beyond the call of duty; and when leaving, to say farewell.

A common format runs through most of the nine chapters: (1) a general, largely undocumented discussion of the subject of the chapter—e.g. discipline; (2) a series of cases illustrating problems actually encountered by student teachers in that area; and (3) broad questions—actually involving the student—which usually follow the cases. Unfortunately, these questions, by focusing largely on solutions rather than on principles underlying proposed solutions,

are not so pointed as the cases. However, this need not deter the instructor who wishes to use these materials, as he may devise his own questions. In addition a brief summary follows each chapter. Sometimes, references for further reading—frequently chapters in standard textbooks—are included.

Students desiring fairly concrete suggestions for solving a wide variety of problems met in student teaching will find them in this volume. These suggestions will be welcomed by many. However, such prescriptions have almost always been reported without their conceptual framework, and the student teacher, unless using this book under the guidance of a knowledgeable instructor, or in a course using one of the more standard texts, will not be given a set of broad principles which will enable problem solving in diverse situations. If used in one of the above supportive contexts, the book would seem exceptionally well suited to its purposes.

REGINALD L. JONES

University of California, Los Angeles
and

L. WARREN NELSON

Miami University (Ohio)

Selected Papers from the Institute for Developmental Studies [of the New York College], Arden House Conference on Pre-School Enrichment of Socially Disadvantaged Children by Martin Deutsch (Chairman). Reprinted from the *Merrill-Palmer Quarterly of Behavior and Development*, July 1964, X, No. 3, 207-309 (+4). [The Merrill-Palmer Institute, 71 East Ferry Ave., Detroit 2, Michigan.] "Introductory Comments," by Martin Deutsch; "The Psychological Basis for Using Pre-School Enrichment as an Antidote for Cultural Deprivation," by J. McVicker Hunt; "Facilitating Development in the Pre-School Child: Social and Psychological Perspectives," by Martin Deutsch; "The Social Context of Language Acquisition," by Vera P. John and Leo S. Goldstein; "Auditory Discrimination and Learning: Social Factors," by Cynthia P. Deutsch; and "Intelligence and Learning," by Martin Whiteman.

In these 103 pages one will find a wealth of theorizing about the cognitive development of children, particularly from ages 1 to 7 and—especially in Cynthia Deutsch's article—some fascinating empirical findings.

In his 40-page article Hunt brilliantly summarizes and interprets the evidence from studies in many disciplines—evidence which convinces him "... that the belief in fixed intelligence is no longer tenable; that development is far from completely predetermined; that what goes on between the ears is much less like the static switchboard of the telephone than it is like the active information pro-

cesses programmed into electronic computers to enable them to solve problems; that experience is the programmer of the human brain-computer, and thus Freud was correct about the importance of the experience which comes before the advent of language; that, nonetheless, Freud was wrong about the nature of the experience which is important, since an opportunity to see and hear a variety of things appears to be more important than the fate of instinctual needs and impulses; and, finally, that learning need not be motivated by painful stimulation, homeostatic need, or the acquired drives based upon these, for there is a kind of intrinsic motivation which is inherent in information processing and action" (pp. 241-242).

Hunt's approach is at least as neurological as psychological, because these are seen to be closely interrelated. He stresses the work of Hebb, Piaget, Freud, and Montessori. Undoubtedly, most knowledgeable readers of his article will differ with him on certain points, but this is the great strength of Hunt's summary, that it provides new orientations and points of departure for further development of the ideas he sets forth clearly but of necessity briefly, some of which depart sharply from current views. His article and Whiteman's have considerable import for older children and adults, too.

Martin Deutsch cites "the necessity for effective cooperation between educators and behavioral scientists, so as to incorporate the growing knowledge of the socio-psychological development of the child into educational procedures in the interests of facilitating realization of his greatest intellectual and social potential" (p. 251). He points out that sound education from the preschool upward is now the chief basis for the trainability and retrainability needed to succeed in today's and tomorrow's complex occupations, and that the necessity for special efforts with children from lower socio-economic-level families is great because they have "the highest proportion of learning disabilities and school dropouts" (p. 251). He goes on to examine the presumed causes of failure of such children to learn effectively in kindergarten through the third grade and concludes that "it would seem that the child from the [special] pre-school and enriched kindergarten classes might best remain in a special ungraded sequence through the third grade level, a period in which he could be saturated with basic skill training, and not be allowed to move on until he has attained basic competence in the skills required by the higher grades" (p. 261).

Martin Deutsch's indictment of the philosophy of most nursery schools is so vivid and pertinent for the other papers that the reviewer cannot refrain from offering Deutsch's own words, rather than trying to summarize the paragraph.

The overgeneralized influence on some sections of early childhood education of the emphasis in the child guidance movement upon protecting the child from stress, creating a supportive environment, and resolving emotional conflicts has done more to

misdirect and retard the fields of child care, guidance, and development than any other single influence. The effect has especially operated to make these fields ineffective in responding to the problems of integrating and educating the non-white urban child. These orientations have conceived the child as being always on the verge of some disease process, and have assigned to themselves the role of protecting the child in the same manner that a zoo-keeper arranges for the survival of his charges. Too frequently a philosophy of protectiveness that asks only about possible dangers has prevailed over any question of potential stimulation of development. The attitude that perhaps helped to create this policy of protectionism can also be seen in the suburban 'mom-ism' that so many sociologists and psychoanalysts have commented on. The child is a far healthier and stronger little organism, with more intrinsic motivation for variegated experience and learning, than the over-protectionists have traditionally given him credit for. (p. 262)

Right as Martin Deutsch probably is about certain excessively tender-minded nursery-school settings, one must not confuse primacy of goals of social and emotional development with physical overprotection, or assume that such goals stifle "intrinsic motivation for variegated experience and learning." A number of nursery schools with which the reviewer has been familiar—especially the Vanderbilt University Cooperative Nursery School, in which he participated as a father during the years 1950-53—firmly rejected the problem-child approach and encouraged extensive child-parent and child-child interactions without being concerned much with the routine physical safety of the children. These two and one-half to six year olds played on jungle gyms, climbed ladders, and sometimes knocked each other down with swings, without evoking tense terror from the supervising adults, who were their mothers. True, the program of such nursery schools was not overtly cognitive, but for non-deprived children it had many of the elements that Martin Deutsch and his collaborators at the Institute for Developmental Studies of the New York Medical College advocate.

John and Goldstein's finding that four-year-old, lower-class Negro children in Manhattan had special difficulty with action words (*digging, tying, pouring, building, picking*), along with other considerations, led them to propose a stability of word-referent vs. amount of corrective feedback two-way grid for studying the difficulty of words "which are abundant in the natural environment of lower-class as well as middle-class children . . ." (p. 270). Also, they consider words as mediators and point out that "antecedent conditions necessary for the development of verbal mediation have not yet been explored" (p. 272). This short chapter contains much provocative material.

Cynthia Deutsch treats auditory recognition and discrimination excellently, and presents rather extensive data of her own that show relationships between several aspects of audition and reading ability. She studied performance of good versus poor readers in several grades across visual and auditory sense modalities and found that "poorer readers have poorer auditory discrimination and . . . greater difficulty in shifting from one modality to the other [than do better readers]" (p. 288). She surmises that ". . . lower-class children, who live in very noisy environments, do not develop the requisite auditory discrimination abilities to learn to read well—or adequately—early in their school careers" (p. 293). This leads her ". . . to postulate that a particular minimum level of auditory discrimination skill is necessary for the acquisition of reading and of general verbal skills" (p. 294).

Whiteman undertakes ingeniously to interrelate the factors of psychometric factor analysis, Harlow's learning sets, and Piaget's operations. He comes to the following conclusion: "Factors are inferred from performance, learning sets from performance and antecedent experience. . . . the presence of an operation should be inferable from the emergence of a factor underlying the performances subsumed under the operation. . . . learning sets may have explanatory value in accounting both for the inter-situational consistency of operations and their developmental aspect" (pp. 300-301).

Then he attempts to set up a systematic hierarchy of factor, operation, and learning set. His exposition is too complex to be summarized briefly here, but one of his statements will be of special interest to measurement specialists: "From a predictive point of view, Piaget's arrangement of cognitive structures forms a massive Guttman scale with formal operations as the highest-order structure, concrete operations next lower down, and intuitive regulations, the pre-conceptual phase, and the sensorimotor stages following in downward progression" (p. 305).

All five papers and Martin Deutsch's brief introduction are excellent contributions to the formulation and measurement of various aspects of "intelligence" and "achievement," the prevention and amelioration of academic retardation, and the quest for keys to education that will make tomorrow's adults readily *retrainable* upward as demands for technical-intellectual competencies change rapidly. This conference report has great implications for the massive "war on poverty" recently launched by Congress. It should be read thoughtfully by "interventionists" who hope to increase the academic and vocational level of culturally disadvantaged persons of all ages—and who is not to some extent disadvantaged?

JULIAN C. STANLEY
Laboratory of Experimental Design
University of Wisconsin, Madison

Reading Disability: Diagnosis and Treatment by Florence Roswell and Gladys Natchez. New York: Basic Books, Inc., 1964. Pp. x + 248. \$5.50.

Here is a book on remedial reading for which instructors of introductory courses at the college level have been waiting. This book is not a patchwork of numerous studies performed by various researchers, under varying conditions, and at various times, beginning with the words, "Research shows . . ." and ending with ". . . the results are therefore equivocal." Such books crammed with studies may be helpful to the uninitiated, but rarely to the practitioner.

This is not to say, however, that the authors disregard the findings of research; rather, as students and as clinicians they have absorbed the various findings, have tried them out, and after years of practice have emerged as professionals who can now prescribe.

Though they prescribe, they inject not an iota of dogmatism; rather, one senses only a deep concern for the "disabled reader" whose relations with peers, family, society, and himself are strained as long as a reading disability remains. And the disability will remain as long as there is a serious discrepancy between reading ability and intellectual potential as assessed by individual intelligence tests and achievement tests.

The book is well organized. The authors cover all the diagnostic skills in the first three chapters by presenting methods of diagnosis on three levels—the level of the classroom teacher, that of the remedial reading teacher, and that of the school psychologist.

In Chapter 4, the authors present some psychotherapeutic principles for establishing a favorable relation between pupil and remedial worker. Then, in the next three chapters, the authors name and describe materials for teaching word recognition, developing vocabularies, improving comprehension, and establishing study skills.

The two chapters that follow deal with special cases outside the normal area. Chapter 8 presents the problems inherent in working with older pupils who are reading at very low levels. More specifically, the problem revolves around finding suitable materials which would appeal to a high school student with a first-grade vocabulary. Chapter 9 deals with the other extreme: the bright high school student who is not achieving up to capacity.

In the final chapter, the authors present six case studies ranging from a brain-damaged boy to an underachiever of high intelligence. The authors use this chapter to demonstrate how to apply the various methods of diagnosis and treatment described throughout the book.

The most valuable of the four appendixes lists the following groups of tests most commonly used in both the classroom and clinic: tests for reading-readiness, silent reading, oral reading, and

diagnostic reading; then tests of intelligence, both group and individual; tests of personality; and tests of a highly specialized nature such as the Harris Tests of Lateral Dominance.

In conclusion, the reviewer believes that the clear and direct language, the simple yet penetrating methods of diagnosis, the appropriate and uncomplicated remedial treatment make the book so readable that any intelligent person could learn much by studying it on his own.

WALTER PAUK
Reading-Study Center
Cornell University

Read with Speed and Precision by Paul D. Leedy. New York: McGraw-Hill Book Company, Inc., 1963. Pp. xiii + 402.

Most manuals on developmental reading for college students are quite similar in both content and organization. This manual, however, is different.

The first difference is in content. The author includes instruction usually found in books intended for introductory courses in college composition and rhetoric. For example, he analyzes the structure of sentences, paragraphs, chapters, and books. Such analysis is good, since the student is given the basis for comprehension through awareness.

The second difference is in organization. The author uses a block-type arrangement; that is, he presents all of the instruction in one continuous, unbroken sequence of 100 pages; practice exercises in another block of 170 pages; and tests of comprehension in another block of 100 pages.

Most other manuals use the alternating method; that is, they present a segment of instruction on a single topic, then several practice exercises; and then sets of questions to test for comprehension. The instruction-exercise-question pattern is thus repeated again and again.

Though the author's block-type organization appears, at first inspection, to have solved the problem of arrangement, the actual use of the manual reveals, however, an inherent weakness.

The inherent weakness is functionality. Notice how involved and cumbersome the process becomes. For example, a selection on page 8 is read; this act necessitates turning to page 279 for the comprehension questions, then to pages 380-381 for the answer key, then to page 389 to the rate conversion chart, then to the inside front cover for the speed chart, then to the inside back cover for the comprehension chart, and finally to page 15 to record the results on a profile chart—all this because one selection was read.

Let one turn, yet, to another weakness: the weakness of presenting a questionable method for gaining speed in reading. The author advocates "scooping up groups of words at a single glance."

Some widely publicized research¹ based on eye-movement photography shows that the average college reader has an average span or intake of only 1.1 words per single eye fixation, and even superior readers seldom take in more than 2.5 words. Thus, the taking in of 1.1 words per "glance" is hardly "scooping up groups of words at a single glance."

The two negative aspects do not, however, affect the quality of the manual. Anyone who inspects this manual will be impressed by the soundness and depth of the instruction, the appropriateness and excellence of the exercises, the genuineness and challenge of the questions. The student who takes to heart the author's instruction, then practices diligently on the exercises provided will emerge, in the reviewer's opinion, reading with speed and precision because now the *mind* can move across the page faster, not merely the eyes.

In conclusion, in the field of college-adult reading, this manual is worthy of a place in the top category.

WALTER PAUK
Reading-Study Center
Cornell University

Psychology (Fourth Edition) by T. L. Engle. New York: Harcourt, Brace and World, Inc., 1964. Pp. 660.

Prepared for high school students, the unit structure of this effectively written book will appeal to most teachers. Since the units are not only broad but also intensive, they should insure a maximum degree of flexibility. The author stresses areas of interest for high school students such as the following titles of chapters suggest: "Emotional Problems of High School Students," "The Growth of Friendship and Love," "Marriage and the Family," "Problems of Society," and "World of Work." These chapters, which are skillfully handled in professional terms, are geared to the comprehensive level of the high school.

Statistical treatment is well conceived in an appendix that could be used as a source of additional instruction or for investigation by the more able students. Normal distribution, correlation, validity, reliability, measures of central tendency, percentiles, deciles, and quartiles are adequately explained.

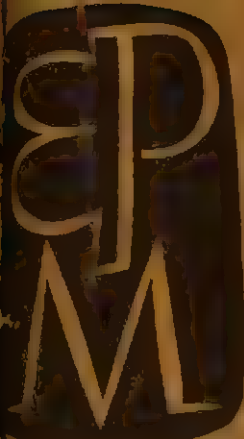
The presentation of the book is orderly, scholarly, and up-to-date. With the exception of being written with a minimum of professional terminology, it could easily pass as a college introductory text.

The detailing and video portions of the book are superb. Professional and timely even to the picture of the astronauts, this text is well worth considering.

ROY M. FITCH
San Fernando Valley State College

¹ Taylor, Stanford E., et al. *EDL Research and Information Bulletin No. 3*, Huntington, N.Y.: Educational Developmental Laboratories, 1960.

3 31 1265



EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

- Alpha Coefficients for Stratified-Parallel Tests.* LEE J. CRONBACH, PETER SCHÖNEMANN, AND DOUGLAS MCKIE 313
- An Empirical Comparison of Methods for Estimating Factor Scores.* JOHN L. HORN 313
- Time-Saving Procedure for Computing Z Scores.* HELEN A. HEATH 323
- Reliability and Validity: Basic Assumptions and Experimental Designs.* EDWARD E. CURETON 327
- The Effects of Partial-Pacing on Test Parameters.* JAMES M. ELLIOTT AND H. G. OSBURN 347
- Risk Taking and Academic Success and Their Relation to an Objective Measure of Achievement Motivation.* ALBERT E. MYERS 355
- 16 PF Item Response Patterns as a Function of Repeated Testing.* KENNETH I. HOWARD AND HERMAN DRESCHLAUS 365
- Interrelationships among MMPI Item Characteristics.* JERRY S. WIGGINS AND LEWIS R. GOLDBERG 381
- Communality and Favorability as Sources of Method Variance in the MMPI.* JERRY S. WIGGINS AND VICTOR R. LOVELL 399

(Continued on inside front cover)

VOLUME TWENTY-FIVE, NUMBER TWO, SUMMER, 1965

<i>Differential Content Validity: The California Spelling Test, an Illustrative Example.</i> KENNETH D. HOPKINS AND CAROLYN J. WILKERSON	413
<i>The EPPS Pattern and the "Nursing Personality."</i> DANIEL V. CAPUTO AND CONSTANCE HANF	421
<i>Factor Analysis of Ranked Educational Objectives: An Approach to Value Orientation.</i> FRED W. OHNMACHT	437
<i>Reading Difficulty of Physics and Chemistry Textbooks.</i> MILTON D. JACOBSON	449
<i>Engineering Freshman Norms for the D.A.T. Mechanical Reasoning and Space Relations Tests Utilizing Fifteen-Minute Time Limits.</i> CHARLES W. JONES AND DAN McMILLEN	459
<i>Social Desirability and the Semantic Differential.</i> LEROY H. FORD, JR. AND MURRAY MEISELS	465
VALIDITY STUDIES SECTION	477
BOOK REVIEWS	619

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Contributors receive one hundred reprints of their articles without charge. Manuscripts should be sent in duplicate to G. Frederic Kuder, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia.

Subscription rate, \$10.00 a year, domestic and foreign. Single copies, \$2.50. Back volumes: Volume V or later, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Editor: G. Frederic Kuder

Associate Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

WILLIAM V. CLEMANS

Science Research Associates, Inc.

LOUIS D. COHEN

University of Florida

HAROLD A. EDGERTON

Performance Research, Incorporated

MAX D. ENGELHART

Chicago City Junior Colleges

E. B. GREENE

Chrysler Corporation

J. P. GUILFORD

University of Southern California

JOHN A. HORNADAY

Houghton Mifflin Company

E. F. LINDQUIST

State University of Iowa

FREDERIC M. LORD

Educational Testing Service

ARDIE LUBIN

U. S. Naval Hospital, San Diego

SAMUEL MESSICK

Educational Testing Service

WILLIAM B. MICHAEL

*University of California,
Santa Barbara*

HOWARD G. MILLER

*North Carolina State University at
Raleigh*

P. J. RULON

Harvard University

C. L. SHARTLE

Ohio State University

KENDON SMITH

*The University of North Carolina
at Greensboro*

THELMA G. THURSTONE

*University of North Carolina at
Chapel Hill*

HERBERT A. TOOPS

Ohio State University

JOHN E. WILLIAMS

Wake Forest College

E. G. WILLIAMSON

University of Minnesota

DOROTHY ADKINS WOOD

*University of North Carolina at
Chapel Hill*

VOLUME TWENTY-FIVE, NUMBER TWO, SUMMER, 1965

ALPHA COEFFICIENTS FOR STRATIFIED-PARALLEL TESTS

LEE J. CRONBACH, PETER SCHÖNEMANN, AND DOUGLAS MCKIE¹

University of Illinois

IN an internal-consistency study investigators almost always rely on the KR20 formula or its generalized version α , a coefficient most appropriate for tests formed by random sampling of items (Cronbach, Rajaratnam, and Gleser, 1963). Several sources (Cronbach, 1951; *Technical Recommendations*, 1954; Lord, 1956; Tryon, 1957) have questioned the suitability of internal-consistency analysis that does not take stratification into account and have recommended instead a modification of the Jackson-Ferguson (1941) "battery reliability" coefficient. But the usefulness of this coefficient, which we shall call α_s , has not been widely appreciated. Rajaratnam, Cronbach, and Gleser (1964) have recently placed the theory of stratified-parallel tests on a more substantial basis, discussing them in the context of a "theory of generalizability." These authors rederive the formula for α_s , present a formula γ_s analogous to KR21, and advocate the application of these formulas to any test constructed by a stratified plan. It is uncertain, however, how much

¹ A study conducted under grant M-1839 from the National Institute of Mental Health. The assistance of Hiroshi Azuma and Kern Dickman is gratefully acknowledged. The current address of Cronbach is School of Education, Stanford University, and that of McKie is School of Education, University of British Columbia.

This appears to be an appropriate place to voice the gratitude of the entire profession for the ILLIAC computing facility during the past ten years. The hospitality of the Digital Computer Laboratory of the University of Illinois to behavioral scientists has had far-reaching benefits. The rush of technological improvement having left her behind, ILLIAC I was retired from service on January 1, 1963; this study is one of the very last of her contributions to psychology.

difference there is between the random-model coefficient and that which takes stratification into account. Nearly all previous empirical studies of internal-consistency formulas have been restricted to randomly parallel tests with uniform content. One limited study (Cronbach, 1951) compared stratified split-half coefficients with coefficients from random splits and found little difference between the two sets of results.

The present study follows in the main the method of the Cronbach-Azuma (1962) study of random-parallel, single-factor tests, and in part replicates it. We specify the statistical properties of a universe of items divisible into subsets representing content categories, direct a computer to construct a series of tests by applying a certain sampling plan, and compute the test reliability by several formulas. We are chiefly concerned with the question: how well does each formula estimate the average squared correlation of the observed test score with the average score on the infinite family of stratified-parallel tests? Our computations range over universes with various statistical structures and over various sampling plans.

Though our report deals with only a limited number of rather special cases, our general theory gives us a basis for confidence in generalizing to cases not studied. Our comparisons therefore give a reasonably good answer to all of the questions of practical importance in test analysis.

Some Theory Regarding Stratified-Parallel Tests

We extract here the theory developed by Rajaratnam, Cronbach, and Gleser (1964). They assume that there is a *universe* of items classified into *strata*. A *sampling plan* dictates that a test is to be formed by drawing, at random, k_h items from each stratum h ; k_h may in principle change from stratum to stratum. The sampling forms a *test* t of k items ($k = \sum k_h$) on which p has a total score X_{pt} whose variance in the population of persons is V_t . All tests that could be formed by applying a particular plan to a particular universe constitute a *family*. Each test t is in effect a random sample drawn from the family. The expectation of the X_{pt} for person p over all tests in the family is denoted by M_p ; this we shall call p 's "family score." Ordinarily there is a different family score for every combination of item universe and sampling plan.

For any specific test t , there is a coefficient of generalizability

ρ_{Mt}^2 , the squared product-moment correlation between test scores X_{pt} and family scores M_p over the population of persons. It reports what proportion of the variance in family ("true") scores is linearly predictable from observed scores. Since tests do not necessarily have equal variances and intercorrelations, this coefficient varies somewhat from test to test; we therefore distinguish between a specific ρ_{Mt}^2 and the expectation $E(\rho_{Mt}^2)$ over all tests in the family, again for the population of persons. As an index of generalizability or reliability for the family we adopt $E(\rho_{Mt}^2)$, which tells how accurately we can expect to generalize from one test to the universe of similar tests. This reflects the degree of equivalence among tests; it of course does not give information about generalization over other facets such as occasions. (Some investigators might choose as an index the correlation between two of the tests, or the expected value over all pairs of tests; for the tests we deal with, $E(\rho_{Mt}^2)$ is certainly very close to $E(\rho_{Mt}^2)$, because the variances of stratified tests of even modest length are close to uniform.) For actual tests and persons, family scores cannot be observed and ρ_{Mt}^2 remains unknown. With hypothetical data, however, we can determine ρ_{Mt}^2 for any specific test, and by averaging coefficients for many tests can determine $E(\rho_{Mt}^2)$ with any desired precision.

The intraclass correlation α_S between tests in the family is an approximation to $E(\rho_{Mt}^2)$; and the estimate of α_S from a single test is close to an unbiased estimate of α_S for the family.

We shall write formulas here in a form corresponding to our computer operations; computational formulas suitable in treating actual data are given by Rajaratnam *et al.* Our formulas and the usual formulas give identical results, save that we assume at all points that we have data from the entire population of persons. We restrict ourselves to sampling plans where k_h is uniform for all m strata, which allows us to simplify certain formulas. Let W_h be a covariance between two particular items in a given stratum, let ΣW_h be the sum of the $k_h(k_h - 1)$ covariances for pairs of items within a subtest drawn from stratum h , and let $\Sigma \Sigma W_h$ be the sum of W_h over all strata. Let U be a covariance between two (unlike) items in different strata, and $\Sigma \Sigma U$ the sum over all such pairs of items in the test.

Various internal-consistency analyses are possible. Consider a universe with two types of content, a and b , and items at two levels

$$\alpha_s = \frac{1}{V_T} \left(\frac{k_h}{k_h - 1} \sum \sum W_h + \sum \sum U \right). \quad (1)$$

		Items							
		1	2	3	4	5	6	7	8
		a A	a A	a B	a B	b A	b A	b B	b B
1	a A		1	2	2	3	3	4	4
2	a A	1		2	2	3	3	4	4
3	a B	2	2		1	4	4	3	3
4	a B	2	2	1		4	4	3	3
5	b A	3	3	4	4		1	2	2
6	b A	3	3	4	4	1		2	2
7	b B	4	4	3	3	2	2		1
8	b B	4	4	3	3	2	2	1	

Figure 1. Schema representing covariances within an 8-item test. a and b are content categories; A and B are difficulty levels. Numerals represent covariances of different types; type 3 covariances, for example, involve items of unlike content but similar difficulty.

of difficulty A and B. Crossing these categories produces four strata, as indicated in Figure 1. Here, an illustrative test has been formed by selecting 2 items at random within each stratum. Thus items 1 and 2 have the same general content a and the same difficulty A. While this test was (hypothetically) constructed by drawing from four strata, it might have arisen from any one of four different sampling plans.

Random sampling. Equivalent to considering all categories together as a single stratum, $k_h = 8$.

Stratification on content alone. Two strata, a and b, 4 items each.

Stratification on difficulty alone. Two strata, A and B, 4 items each.

Stratification on content and difficulty. Four strata, 2 items each. Each of these modes of test construction would generate a different family, any of which the present test might belong to. For each family there is a different $E(\rho_{MF}^2)$, with the finer stratifications producing the higher values. If the given test is analyzed by formula (1), different values of α_s will be obtained when different stratifications are made. Analysis may be based on any of the plans stated

above; each leads to a coefficient which the theory says is an estimate of $E(\rho_{Mt}^2)$ for the family defined by that plan. We distinguish these four values by designating them α , α_G , α_D , and α_{GD} , respectively.

If the analysis treats all items as belonging to a single stratum, all covariances represented in Figure 1 are "within-stratum," and (1) gives a result identical to that from the α (KR20) formula. This analysis has the effect of averaging all covariances of types 1, 2, 3, and 4, entering this average in the 8 diagonal cells, and dividing the total of the 64 entries by the test variance. According to theory, α approximates $E(\rho_{Mt}^2)$ for the family of randomly parallel tests. Three empirical questions arise. The first—how good is the approximation for randomly parallel tests formed from items having a single common factor?—was answered in the Cronbach-Azuma study. Only in certain unlikely situations is the correspondence unsatisfactory. The second—how good is the approximation for randomly sampled tests where items cluster around various common factors?—has not been studied. The third—how good is the approximation when the family score is that for a family of stratified-parallel tests?—is examined in this paper. This amounts to asking how much is lost when one obtains α instead of the α_S that is theoretically appropriate for the stratified family.

If the analysis takes both content and difficulty into account, covariances of type 1 are W 's and all others are U 's. Formula (1) in effect averages the type-1 covariances within each stratum and enters that average in the diagonal. Summing the matrix and dividing by V_t gives α_{GD} . If content strata only are to be taken into account, covariances of types 1 and 2 are W 's; the resulting coefficient is α_G . In α_D the diagonal entries are determined by averaging the type-1 and type-3 covariances. This analysis recognizes difficulty strata but not content strata.

There is a series of " γ " formulas that provide lower bounds to the several α 's, just as KR21 provides a lower bound to the KR20 coefficient (Rajaratnam *et al.*, 1964). We have calculated γ values for selected tests treated below, but we find that no simple generalization can be made. To discuss the γ values adequately would overburden the present paper. The stratified γ underestimates the corresponding stratified α to such an extent that it cannot be used under many circumstances. But as k_s increases or as the within-

stratum range of P_i decreases, γ becomes a good estimator of the corresponding α . We hope to present more definite results on γ in a subsequent paper.

The dichotomous items (scored 1 or 0) that many tests employ have properties that have caused some concern to test theorists. In the Kuder-Richardson development of the α formula it was assumed that the item intercorrelation matrix had unit rank. Since a matrix of phi coefficients for items differing in difficulty cannot have unit rank, this implied that all items were of uniform difficulty. There have been various ways of circumventing this problem. The derivation by Rajaratnam *et al.* simply argues that the expected α_S is a lower bound, and an approximation, to $E(\rho_{Mt}^2)$. If items can be classified into strictly homogeneous content strata, then as stratification on difficulty becomes increasingly fine the value of α_S approaches the value of $E(\rho_{Mt}^2)$ for the family so defined. But fine stratification is rare in practice. A significant question, then, is how adequately a coarse stratification on difficulty reduces the discrepancy between α_S and $E(\rho_{Mt}^2)$.

One important theoretical point remains to be made. In a certain sense, α_S obeys the Spearman-Brown rule. To be precise, consider two families of tests, of different lengths but based on proportional sampling plans (plans such that the k_h of one test are proportional to the respective k_h of the other). Then the expected value of α_S over the family of longer tests is related to the expected value for the family of shorter tests according to the Spearman-Brown formula, provided that the stratification in analysis coincides with that used in selecting items. As a consequence, we can restrict our empirical study to tests of a single length, and yet draw conclusions regarding the behavior of these coefficients for longer and shorter tests.

We would like to draw conclusions about the discrepancy between α_S and $E(\rho_{Mt}^2)$ for any length of test. Though we have no theoretical basis for extrapolating $E(\rho_{Mt}^2)$ to other test lengths, there is empirical evidence that $E(\rho_{Mt}^2)$ for randomly parallel tests comes very close to following the Spearman-Brown formula. Table 1 shows values of $E(\rho_{Mt}^2)$ for single-factor randomly parallel tests with 9, 18, and 54 items having various levels of interitem tetrachoric correlations r_w . (The data for the 18-item tests were collected during the present study; the remainder are taken from Cronbach and Azuma.) Here and elsewhere, we shall find it illuminating to transform correla-

TABLE 1
Correspondence of S/N Equivalents for $E(\rho_M^2)$ to Values Consistent with the Spearman-Brown Formula (Single-Factor Tests Only)

Interitem correlation r_w	.30				.70				1.00			
	9	18	54	9	18	54	9	18	54	9	18	54
Test length												
$E(\rho_M^2)$.613	.757	.903	.839	.913	.969	.920	.961	.986			
S/N equivalent	1.58	3.12	9.31	5.21	10.49	31.26	11.50	24.64	70.43			
Ratio of S/N values	1	2.0	5.9	1	2.1	6.0	1	2.1	6.1			

If the Spearman-Brown formula were valid for $E(\rho^2)$, the ratio of S/N values would be 1:2:3.

tional indices such as α into "signal-noise ratios" of the form $\alpha/(1 - \alpha)$; for a discussion of this index, see Cronbach and Gleser (1964). Under the hypothesis that $E(\rho_{xt}^2)$ obeys the Spearman-Brown formula, the signal-noise equivalents for the three lengths of test should exhibit the ratio 1:2:6; for each r_w , the observed ratio is close to the expectation. Because the variances of stratified tests are more nearly uniform than those of randomly parallel tests, we expect $E(\rho^2)$ for stratified tests to conform even more closely to the Spearman-Brown prediction. No data for heterogeneous tests are available, however.

Although, for a test stratified in a given manner, the average α_s obeys the Spearman-Brown rule if the analysis uses the same stratification as did the test construction, this is not the case if some other stratification is used in analysis. Particularly, if tests are constructed on the basis of content-and-difficulty stratification, the Spearman-Brown rule does not hold for α , α_C , or α_D ; only α_{CD} for those strata obeys the rule. A formula describing the change in α with change in length of a stratified test is given by Rajaratnam *et al.* The striking consequence of this formula is that where two coefficients are computed, one for the sampling plan used in test construction and one for some coarser stratification, the ratio of the corresponding S/N values approaches a limit as k increases. The limit is expressible in terms of two average covariances— W , the average covariance for items within the same stratum of both plans, and W' , the average for items within the same stratum of the coarser plan only—and the item variances. We shall give the formula only for the case where one plan uses three strata (k , uniform) that are further divided into thirds in the finer plan. Write v for the average item variance. Then

$$\text{Lim } \frac{S/N \text{ (coarse)}}{S/N \text{ (fine)}} = \frac{3v - 3W}{3v - W - 2W'} \quad (2)$$

Procedures

A hypothetical universe of items is specified; this is divided into three content categories (except in one substudy where six are used). Items within a category have a uniform tetrachoric correlation r_w , and a specified correlation r_b with each item in another category. While it is not necessary to make r_w the same for all categories or

r_b , the same for all pairs of categories, in our computations these are kept uniform. While there are many ways of specifying universe characteristics other than those we employ (for example, uniform phi correlations or uniform covariances might easily have been introduced), our specifications appear to cover much of the range of test types.

	a	b	c		a	b	c		a	b	c
a	.30	.30	.30	a	.50	.30	.30	a	.50	.00	.00
b	.30	.30	.30	b	.30	.50	.30	b	.00	.50	.00
c	.30	.30	.30	c	.30	.30	.50	c	.00	.00	.50
	(1)				(2)				(3)		

Figure 2. Specimen tetrachoric correlations among items representing three content categories.

The representative values of r_b and r_w in Figure 2 indicate the flexibility of our model. In each matrix, the diagonal entries are r_w and the off-diagonal entries are r_b . The first set of correlations implies that there is just one common content factor among the test items. In matrix (2) a fairly strong general factor links items of the three types, and slightly less influential factors link items within categories. Matrix (3) exhibits three orthogonal content factors.

All tests under study consist of dichotomously scored items. Item difficulties in the universe are specified by defining a range of P_i and a rectangular distribution within that range. Two ranges are used: the "limited range" $.60 \leq P_i \leq .99$ (comparable to many ability tests), and the "wide range" $.01 \leq P_i \leq .99$. The latter is unrealistically wide, but it thereby provides a severe test of the susceptibility of our formulas to differences in item difficulty. For stratification on difficulty, the range is divided into segments: .60-.79 and .80-.99; or .01-.30, .31-.69, and .70-.99.

The sampling plans are arranged so that both content and difficulty or either one alone can be taken into account. The computer program allows the use of from one to nine strata. To each stratum is assigned one type of content (i.e., the stratum is identified with one column of an r_b - r_w matrix), a value of k_n , and a "difficulty segment." We restrict our studies to sampling plans where difficulty segments and content categories are completely crossed; that is, we do not draw difficult items from one content category and easier items from another.

The universe specification and the sampling plan together define a family of tests. Drawing a particular set of P_i from a stratum by means of random numbers to conform to the sampling plan defines a subtest. The several subtests form a test. Once the P_i for the test are selected, the computer considers in turn each pair of items, enters their parameters (P_i , P_j , and r_w or r_b) in the usual series approximation to calculate their product-moment covariance C_{ij} , and stores it according to its type, as defined in Figure 1. The V_i are also calculated. Then, by appropriately cumulating covariances and variances, the computer determines V_T and α , α_G , α_D , and α_{GD} . This calculation is repeated with the same P_i and a new intercorrelation matrix, to generate a test from another family. After this step has been repeated for each correlation matrix under consideration, a new set of P_i is drawn according to the same sampling plan and coefficients for the new test are calculated, using each correlation matrix in turn. To conserve computer time, only a few tests of each family were constructed and analyzed; the number of tests sampled for any family was selected so as to reduce the standard error of the mean coefficient to an acceptable level.

It remains to describe the determination of $\rho_{M_i}^2$, whose expected value serves as the "ideal" coefficient in our theory. It can be shown that when all k_h are equal and rectangular distributions of P_i are assumed,

$$\rho_{M_i}^2 = \frac{\left(\sum_h \sum_{i=1}^{h_h} \sum_{h'} C_{i_h M_h} \right)^2}{V_i \cdot \sum_h \sum_{h'} C_{M_h M_{h'}}} \quad (3)$$

where M_h (or more properly M_{ph}) is the expected value of a person's scores on subtests drawn from stratum H . Here, h may be the same as h' . To evaluate this, it is necessary to compute* the $C_{i_h M_h}$ and $C_{M_h M_{h'}}$. The computer generates a complete matrix of covariances for each tetrachoric correlation under study and all pairs P_i , P_j in the range .01 - .99. For any item i_h , $C_{i_h M_h}$ is determined by identifying the tetrachoric correlation between items in stratum h' and the stratum to which i_h belongs (which may or may not be h'), identifying the difficulty segment assigned to stratum h' and averaging the C_{ij} for the specified r and P_i over all P_j within that segment. To determine

* Actually, the procedure described here gives $C_{i_h M_h}/k_h$ and $C_{M_h M_{h'}}/k_h^2$; but the k_h cancel out and can be ignored.

$C_{M_i M_j}$, it is necessary to average C_{ij} over the segment of P_i assigned to h and the segment of P_j assigned to h' . There is a different $\rho_{M_i}^2$ for each sampling plan and set of r_b and r_w , but the amount of calculation is greatly reduced by recognizing that many steps of the computation are the same for different sampling plans. For any one sampling plan, r_b , and r_w , several tests were formed and $\rho_{M_i}^2$ for these tests were averaged to obtain an estimate of $E(\rho_{M_i}^2)$ for the family.

A by-product of our work was a set of tables reporting $\bar{C}_{h_i h_j}$ for any two strata. There is one table for each r_w , giving a mean covariance for any pair of difficulty segments of width .10. From these tables one can calculate $E(C_{ii})$ and $E(V_i)$ for tests drawn according to virtually any sampling plan. The ratio of these two is very close to the expected value of unstratified α , which is $E(C_{ii}/V_i)$. By appropriate separation of within- and between-stratum covariances one can also calculate α_s for any stratification. We have used this method in certain subordinate analyses and checks. We believe that questions will arise in the future where our tables will serve other investigators. Moreover, they may be helpful in instruction, since students can design tests with different hypothetical specifications and compute various internal-consistency coefficients and so learn more about the properties of dichotomous items. Copies of the tables together with directions for computing the several α and γ coefficients will be supplied upon request to the Bureau of Educational Research, University of Illinois, Urbana, Illinois.

Results for Single-Factor Tests

Unstratified Tests. We consider first a set of 18-item single-factor tests like the 9- and 54-item tests examined by Cronbach and Azuma. Though three content-strata are formally represented, setting r_b equal to r_w produces items of uniform content so that the stratification "on content" has no meaning. For each of the pseudo-strata, six items were drawn from the difficulty range .01-.99.

Table 2 presents the average values of $\rho_{M_i}^2$ and α for several tests of each family. The mean $\rho_{M_i}^2$ is our best estimate of $E(\rho^2)$; for convenience this mean is referred to throughout as $E(\rho^2)$. α_D , being identical to α for these tests, is not reported. α_O also is omitted since, for a single-factor test, α_O and α differ only by chance; of the two, α is the more dependable because the mean covariance "entered in the diagonal" is based on a greater number of sampled covariances.

TABLE 2

Data for Single-Factor Randomly Parallel Tests with Wide P Range

r_w	ρ_{MT}^2	α	E(C)
			E(V)
1.00	.961 \pm .004	.948 \pm .003	.947
	(.94-.98)	(.93-.96)	
	<i>24.6</i>	<i>18.2</i>	
.70	.913 \pm .002	.903 \pm .003	.904
	(.91-.92)	(.89-.92)	
	<i>10.5</i>	<i>9.31</i>	
.50	.860 \pm .003	.851 \pm .004	.853
	(.85-.88)	(.83-.87)	
	<i>6.14</i>	<i>5.71</i>	
.30	.757 \pm .005	.751 \pm .005	.753
	(.74-.79)	(.73-.78)	
	<i>3.12</i>	<i>3.02</i>	

Each entry for $E(\rho^2)$ and α gives the mean, s.e.M., and range for ten tests. Italicized figures are S/N equivalents of the corresponding means.

Although it is customary to compare correlational indices directly, we have found this somewhat misleading because numerically small differences between very high coefficients may have much practical significance. Transforming coefficients into "signal-noise ratios" often gives a better basis for interpretation (Cronbach and Gleser, 1964). For example, the averages for ρ^2 and α where $r_w = 1.00$, .961 and .948, seem close together—but the corresponding S/N ratios are 25 and 18. A test with S/N 18 must be lengthened about 40 percent to attain an S/N of 25. A coefficient of .961 therefore implies that the test is 140 percent as efficient as the coefficient of .948 implies. Only where the S/N equivalents are close together are the implications of the two coefficients the same. To assist in interpreting results, all average coefficients have been converted to S/N; these conversions appear in italics in our tables. As a rule of thumb, we accept one coefficient as an adequate estimator of another if their S/N equivalents have the ratio .83:1.00 or 1.00:1.20.

For homogeneous tests, $E(\alpha)$ is close to $E(\rho^2)$ even when the tetrachoric correlation is as high as .70. The reader is reminded that interitem tetrachoric correlations for actual tests are in the neighborhood of .05-.40, rarely higher. This confirms the Cronbach-Azuma conclusion that for tests constructed by random sampling from homogeneous items, α provides, on the average, a very good estimate of $E(\rho^2)$. But the range of individual α 's is noteworthy; for $r_w =$

TABLE 3
Single-Factor Tests Stratified on Difficulty: Mean, S.E., and Range for Each Coefficient

r_b and r_w	Wide Range				Limited Range			
	ρ_{Mt}^2	σ_D	α	ρ_{Mt}^2	σ_D	α	ρ_{Mt}^2	α
1.00	.978 \pm .003 (.96-.99) 44.5	.970 \pm .002 (.96-.98) 22.3	.942 \pm .002 (.93-.95) 16.2	.983 \pm .003 (.97-.996) 57.3	.982 \pm .001 (.98-.99) 54.6	.967 \pm .002 (.96-.97) 29.3	.983 \pm .003 (.97-.996) 57.3	.967 \pm .002 (.96-.97) 29.3
.70	.919 \pm .001 (.91-.92) 11.3	.915 \pm .003 (.90-.92) 10.8	.902 \pm .003 (.89-.91) 9.20	.921 \pm .004 (.89-.93) 11.7	.923 \pm .001 (.92-.93) 12.0	.915 \pm .002 (.90-.92) 10.8	.921 \pm .004 (.89-.93) 11.7	.915 \pm .002 (.90-.92) 10.8
.50	.863 \pm .003 (.86-.87) 6.50	.859 \pm .003 (.85-.87) 6.09	.851 \pm .003 (.83-.86) 5.71	.866 \pm .003 (.86-.87) 6.46	.863 \pm .002 (.86-.87) 6.30	.856 \pm .003 (.84-.87) 5.94	.866 \pm .003 (.86-.87) 6.46	.856 \pm .003 (.84-.87) 5.94
.30	.760 \pm .003 (.75-.77) 3.17	.754 \pm .004 (.74-.77) 3.07	.752 \pm .005 (.73-.76) 3.03	.755 \pm .003 (.75-.76) 3.08	.753 \pm .003 (.74-.77) 3.05	.747 \pm .004 (.73-.76) 2.95	.755 \pm .003 (.75-.76) 3.08	.747 \pm .004 (.73-.76) 2.95

Wide-range averages for ρ_{Mt}^2 , σ_D and α are based on 10 tests when $r_w = 1.00$. Limited-range averages for ρ_{Mt}^2 , σ_D and α are based on 10 tests when $r_w = 1.00$ or .70. All other averages are based on 5 tests.

.30, the S/N range is 2.7 to 3.5 (corresponding to α 's of .73 and .78). In longer tests, the range of α would be less.

Cureton (1958) criticized derivations of internal-consistency formulas. These formulas invariably involve the ratio of estimates of $E(C_{ii})$ and $E(V_i)$, even though the theory calls for obtaining the expected value of $E(C_{ii}/V_i)$. Working from our tables we calculated $E(C_{ii})/E(V_i)$; this value is entered in Table 2 under the heading $E(C)/E(V)$. Comparing this to our sample mean of α , which is an unbiased estimate of $E(C/V)$, we see that for each r_w the two values agree within .002. Discrepancies between $E(C)/E(V)$ and α_S were no larger than .01 for the stratified-parallel families studied. Our results therefore seem to dispose of Cureton's concern. At least for tests of 18 or more dichotomous items, the ratio of expectations of C_{ii} and V_i agrees excellently with the expectation of the ratio and the usual substitution of the former for the latter is acceptable.

Homogeneous Tests Stratified on Difficulty. Specifying a certain number of items per difficulty stratum constrains the distributions of P -values in the test. We expected $E(\rho^2)$ for families of tests stratified on difficulty to be somewhat higher than for random-parallel tests made up from the same universe of items, because the former tests are more perfectly equivalent.

We constructed two series of tests. In the wide-range series, six items were selected from each of the three difficulty segments: .01-.30, .31-.69, and .70-.99; in the "limited range" series, nine items were drawn from the segments .60-.79 and .80-.99. Again, three content strata were formally represented, but $r_w = r_b$ so that content is homogeneous. Table 3 shows the average coefficients for both series together with their ranges, standard errors, and S/N equivalents. Again, α_O differs from α only by chance, and α_{OD} from α_D .

The tests of the wide-range series are just like those of Table 2 save for stratification on difficulty. $E(\rho^2)$ shows some increase from difficulty stratification, not very important unless $r_b > .70$ (cf. S/N values). $E(\alpha)$ has changed negligibly, except for a decrease at $r_w = 1.00$. The ranges of all coefficients are reduced by difficulty stratification.

When we compare each estimator to $E(\rho^2)$ for either wide- or narrow-range tests, we find that α_D is, as expected, a good estimator except for the wide-range test with $r_w = 1.00$. In general, α is only slightly less satisfactory as an estimator, and since it is easier to

TABLE 4
Three-Factor Tests Stratified on Content and Difficulty: Mean, S.E.M., and Range for Each Coefficient

r_w	Wide Range					Limited Range				
	ρ_{TM}^2	α_{CD}	α_C	γ_{CD}	ρ_{TM}^2	α_{CD}	α_C	γ_{CD}		
1.00	.926 \pm .004 (.90-.94) 12.5	.912 \pm .006 (.87-.93) 10.4	.825 \pm .006 (.79-.85) 4.71	.899 \pm .011 (.86-.92) 8.90	.944 \pm .006 (.92-.97) 16.9	.949 \pm .004 (.93-.97) 18.6	.899 \pm .004 (.87-.93) 8.90	.946 \pm .008 (.92-.96) 17.5		
.70	.787 \pm .005 (.77-.80) 3.69	.780 \pm .005 (.75-.79) 3.55	.744 \pm .007 (.71-.76) 2.91	.758 \pm .012 (.73-.78) 3.13	.797 \pm .003 (.77-.81) 3.93	.800 \pm .003 (.78-.82) 4.00	.776 \pm .005 (.75-.80) 3.46	.785 \pm .008 (.76-.81) 3.66		
.50	.675 \pm .005 (.66-.69) 2.08	.669 \pm .005 (.64-.68) 2.02	.647 \pm .007 (.61-.66) 1.83	.643 \pm .008 (.62-.67) 1.80	.680 \pm .003 (.67-.69) 2.13	.677 \pm .005 (.66-.70) 2.10	.659 \pm .007 (.63-.69) 1.93	.662 \pm .009 (.64-.69) 1.96		
.30	.512 \pm .005 (.50-.52) 1.05	.508 \pm .005 (.48-.52) 1.03	.496 \pm .007 (.46-.51) .98	.475 \pm .009 (.44-.50) .90	.505 \pm .003 (.49-.51) 1.02	.503 \pm .005 (.49-.52) 1.01	.491 \pm .006 (.47-.52) .96	.483 \pm .011 (.45-.51) .93		

For $r_w = 1.00$, means of ρ_{TM}^2 and the α -coefficients are based on 10 tests. All other means are based on 5 tests.

compute it may be advantageous unless a test has remarkably high item intercorrelations.

Results for Heterogeneous Tests

Tests Stratified on Content and Difficulty. The next set of tests to be considered is formed from a universe with three orthogonal content categories ($r_b = 0$). With a wide range of difficulty there are three content categories crossed with three difficulty segments; in the limited-range universe we cross the content categories with two difficulty segments. These tests have much lower coefficients than those of Tables 2 and 3; using three orthogonal content categories we get coefficients similar to those for a six-item single-factor test having the same r_w .

While it is possible to analyze these tests by formulas that ignore content stratification, this is distinctly inadvisable. This is shown by results for limited-range tests where $r_w = .50$; the S/N equivalents are 2.1 for $E(\rho^2)$, 1.4 for α_D , and 1.4 for α . The results are only a trifle better when $r_w = .30$, and much worse for higher r_w . Analyzing a short heterogeneous test by the simple α formula gives misleading results.

The formulas that take content stratification into account give the results in Table 4. α_{CD} is an excellent estimator of $E(\rho^2)$. While α_C gives somewhat lower estimates, it is close enough to $E(\rho^2)$ to be practically useful when r_w is in the normal range.

Tests Stratified on Content Only. The final major series of tests

TABLE 5
*Three-Factor Wide-Range Tests Stratified on Content Only:
Mean, S.E.M., and Range for Each Coefficient*

r_w	r_b	ρ^2_{MI}	α_C	α
1.00	0	.866 \pm .015 (.77-.93) 6.46	.860 \pm .009 (.82-.91) 6.14	.758 \pm .008 (.73-.80) 3.13
.70	0	.765 \pm .008 (.72-.80) 3.26	.757 \pm .007 (.73-.80) 3.12	.668 \pm .006 (.64-.70) 2.01
.50	0	.664 \pm .007 (.63-.70) 1.98	.656 \pm .007 (.63-.70) 1.91	.579 \pm .006 (.56-.62) 1.38
.30	0	.507 \pm .007 (.48-.55) 1.03	.501 \pm .006 (.47-.54) .965	.442 \pm .006 (.42-.48) .792

was constructed by stratifying on content only, using a wide range of item difficulties. Table 5 shows that α_c is, on the average, extremely close to $\rho_{\mu t}^2$, performing considerably better than α . As in the single-factor case of Table 2, coefficients from individual tests within a family vary greatly. Variation among coefficients will decrease rapidly as the number of items increases or the range of P_i is reduced.

Trend with Change in r_b . We have so far compared the limiting cases where $r_b = 0$ and $r_b = r_w$. In actual tests, r_b is likely to fall between these extremes. We have made only a limited study of

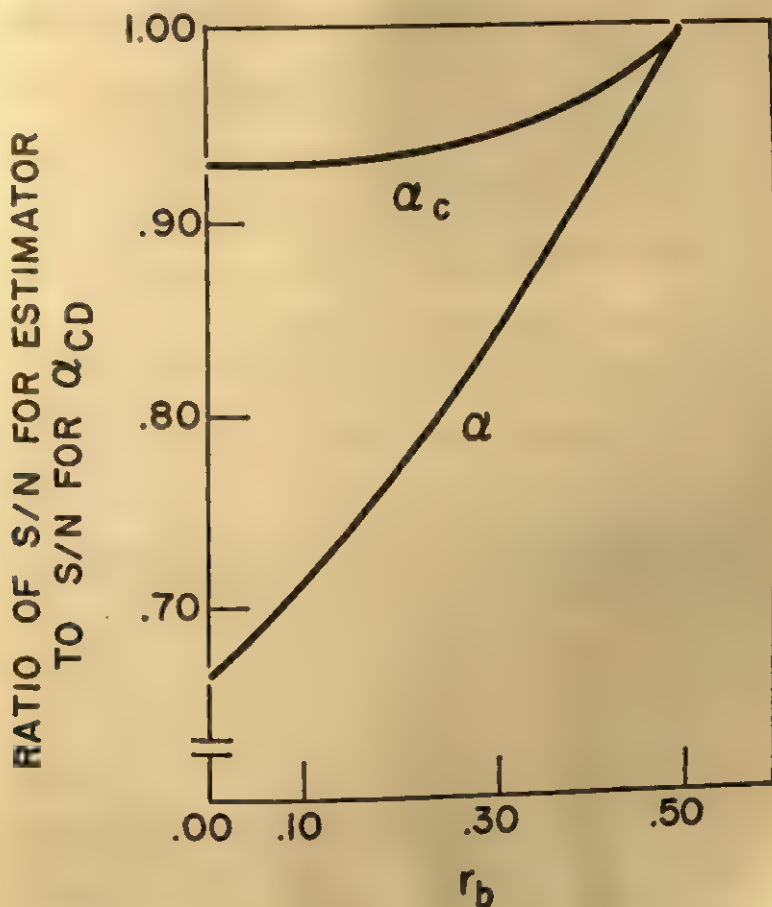


Figure 3. Adequacy of α and α_c as a function of between-stratum item correlations ($r_w = .50$, limited range).

intermediate cases, but these are in sufficiently close agreement that only one set of illustrative results need be reported. For limited-range tests with $r_w = .50$ and $r_b = .00, .30$, or $.50$, the tables of covariances were used to estimate $E(\alpha_{OD})$, $E(\alpha_O)$, and $E(\alpha)$. Since we have previously established that α_{OD} is a fairly good approximation to $E(\rho^2)$, we use it as the standard of comparison to avoid computational labor. Figure 3 shows that the relation of α_O to α_{OD} is very much the same for all r_w , but that α becomes a considerably poorer estimator as r_b decreases. Similar results are found with wide-range tests and with $r_w > .50$. The goodness of any estimator is to be judged from our earlier tables when $r_b = .00$ or $r_b = r_w$. For intermediate r_b , linear interpolation between these values will indicate the approximate size of the estimator.

Effect of Increasing k_h . We expect all coefficients to approach 1.00 as tests are lengthened by increasing k_h proportionately for all strata. There are grounds for expecting the S/N values for $E(\rho^2)$ and α_S to maintain the same ratio as k_h changes, provided that α_S is based on the plan used in test construction. We wish to know also how the relations among α , α_O , and α_{OD} change as a test is lengthened.

From our tables of covariances and formula (2) we calculate the limit of the ratio of S/N values for wide-range tests, presented in Table 6. An additional ratio is available for $r_w = .50$, $r_b = .30$ —.869:.941:1.00. For long tests, it appears that α_O is a satisfactory

TABLE 6

Limiting Values of Ratios $SN(\alpha) : SN(\alpha_O) : SN(\alpha_{OD})$ Corresponding to Stratified Alpha with Coarse and Fine Stratification (Wide-Range Tests)

r_w	$r_b = 0$	$r_b = r_w$
1.00	.338 : .551 : 1	.551 : .551 : 1
.70	.652 : .870 : 1	.870 : .870 : 1
.50	.780 : .941 : 1	.941 : .941 : 1
.30	.881 : .977 : 1	.977 : .977 : 1

substitute for α_{OD} when r_w is .80 or below. For 18-item tests (wide range) α_O is satisfactory if $r_w < .70$. We conclude that for real tests of any length α_O is acceptable in place of α_{OD} , which is harder to compute. Moreover, for homogeneous tests α is acceptable in place of α_D (which equals α_{OD}).

When α is compared with α_G or α_{GD} for very long tests where $r_b = 0$, we find α acceptable so long as r_w is less than about .40. But α is quite unsatisfactory for an 18-item test. The minimum length where α may be used in place of α_{GD} (by the .83 criterion) is 54 items when $r_w = .30$ and $r_b = 0$. The discrepancy between α and α_G decreases as r_b increases, the trend in the S/N ratio being nearly linear. The discrepancy is slightly lessened when the difficulty range is narrowed.

Summary and Recommendations

To explore the properties of various internal-consistency formulas, we have analyzed hypothetical stratified-parallel tests constructed by sampling items from universes with specified characteristics. Formulas generally thought of as approximations or lower bounds to the squared correlation of test score with true test score ($E(\rho_{Mt}^2)$) were compared to a direct numerical estimate of the latter value, so that the goodness of approximation can be judged. We compare several " α " coefficients, versions of the stratified intraclass correlation whose degenerate case with a single stratum is the α or KR20 formula.

While we have dealt with only a few of the possible universe specifications and sampling plans, the reader can extend the findings by interpolation and extrapolation. All tests in this study are 18 items long. It is known that as test length changes the stratified α coefficients calculated on the basis of the stratification used in test construction obey the Spearman-Brown rule. There is some evidence that $E(\rho^2)$ follows that rule fairly closely. Homogeneous tests draw items from a single content stratum whose degree of homogeneity is defined by the tetrachoric interitem correlation r_w . Our tests draw items from three content strata. The interitem tetrachoric correlations are labelled r_w for items within a stratum and r_b for items in different strata. In most of our analyses, $r_b = r_w$ (single-factor test) or $r_b = 0$.

The findings are too complex to be compactly summarized, but a single table of particularly representative values will present the heart of the results. Computations were made assuming r_w of .30, .50, .70, and 1.00. While the anomalous behavior of product-moment correlations at high tetrachoric levels makes the higher r_w interesting, it is the lower levels that are encountered in actual tests. We re-

produce in Table 7 results only for within-stratum correlations of .50 (adding, for completeness, a few values not presented earlier).

TABLE 7
Summary of Results for 18-item Tests with $r_w = .50$

Table	Universe Content	Difficulty Range	Stratification in sampling	S/N Corresponding to Mean Coefficient				
				ρ^2_{Mt}	α_{CD}	α_C	α_D	α
2	One factor	Wide	None	6.1				5.7*
3	One factor	Wide	Difficulty	6.3			6.1*	5.7*
3	One factor	Limited	Difficulty	6.5			6.3*	5.9*
4	One factor	Wide	Content and difficulty	2.1	2.0*	1.8*	1.3	1.3
4	Three factors	Limited	Content and difficulty	2.1	2.1*	1.9*	1.4	1.4
5	Three factors	Wide	Content	2.0		1.9*		1.4

*Acceptable as estimate of $E(\rho^2)$.

Values in boldface are results for the coefficient theoretically appropriate to tests stratified as indicated.

All coefficients have been converted to equivalent signal-noise ratios, which are directly related to an investigator's decisions about whether a certain test length is appropriate for the degree of generalizability he requires. As a rule of thumb, we suggest that an estimate is seriously misleading if S/N for the estimate is less than .83 times the S/N corresponding to the parameter value. Any two S/N values will retain roughly the same ratio to each other as tests are lengthened, other things being equal. For realistic tests—where r_w is likely to fall below .50—the discrepancies among $E(\rho^2)$ and the coefficients will ordinarily be less than those in this table. The reader should bear in mind that we have investigated extremely heterogeneous tests; the typical test has correlated content strata, whose behavior will fall between that of the one-factor and three-factor tests of this table.

The results are clear. When a test is constructed by stratifying on *content and difficulty*, one may properly estimate its coefficient of generalizability by α_{CD} or α_C . The latter, though less precise, is generally easier to compute. For a test stratified on *content*, α_C should be used. With r_w below .20, one may use the simpler α formula. For a single-factor test stratified on *difficulty* only, the best estimator is α_D , but the more easily computed α gives acceptable results

unless strata are extremely narrow and r_w is high. For *unstratified* (random-parallel tests), α gives acceptable results.

Stratifying on content is clearly more important than stratification on difficulty, both in test construction and test analysis. The so-called difficulty factors that have received so much attention from some test theorists prove to have very little influence on α coefficients unless r_w is unrealistically high.

This paper has examined tests for which an explicit sampling plan or table of specifications was laid down during test construction. Nothing in our procedure, however, makes the results inapplicable to a *posteriori* stratification, where the investigator sorts items from an existing test into strata of his own defining, and estimates the coefficient of generalizability over the family defined by this postulated sampling plan.

In sum, our results strongly support the *Technical Recommendations*: "If a test can be divided into sets of items of different content, internal consistency should be determined by procedures designed for such tests"—always assuming that the intention is to generalize over other tests covering these same content categories.

REFERENCES

- Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, XVI (1951), 297-334.
- Cronbach, L. J. and Azuma, Hiroshi. "Internal-consistency Reliability Formulas Applied to Randomly-sampled Single-factor Tests: An Empirical Comparison." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 645-665.
- Cronbach, L. J. and Gleser, Goldine C. "The Signal-Noise Ratio in the Comparison of Reliability Coefficients." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIII (1964), 467-480.
- Cronbach, L. J., Rajaratnam, Nageswari, and Gleser, Goldine C. "Theory of Generalizability: A Liberalization of Reliability Theory." *British Journal of Statistical Psychology*, XVI (1963), 137-163.
- Cureton, E. E. "The Definition and Estimation of Test Reliability." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVIII (1958), 715-738.
- Jackson, R. W. B. and Ferguson, G. A. *Studies on the Reliability of Tests*. Bulletin No. 12. Department of Educational Research, University of Toronto, 1941.
- Lord, F. M. "Sampling Error Due to Choice of Split in Split-Half Reliability Coefficients." *Journal of Experimental Education*, XXIV (1956), 245-249.
- Rajaratnam, Nageswari, Cronbach, L. J., and Gleser, Goldine C.

"Generalizability of Stratified-Parallel Tests." *Psychometrika*, 1965, in press.

"Technical Recommendations for Psychological Tests and Diagnostic Techniques." Washington, D. C.: American Psychological Association, 1954. (*Psychological Bulletin*, L (1954), Supp.)

Tryon, R. C. "Reliability and Behavior Domain Validity: Reformulation and Historical Critique." *Psychological Bulletin*, LIV (1957), 229-249.

AN EMPIRICAL COMPARISON OF METHODS FOR ESTIMATING FACTOR SCORES

JOHN L. HORN
University of Denver

METHODS which can be used to estimate factor and component scores have been described in a number of publications (Baggaley and Cattell, 1956; Bartlett, 1937; Harman, 1941, 1960; Horn, 1964; Horst, 1941; Kaiser, 1962; Kestelman, 1952; Ledermann, 1939; Thomson, 1949; Thurstone, 1947). Interest in these methods has been largely academic, however. To date few studies have used any of the available methods in actual research and fewer still have subjected the methods to empirical comparisons, the Baggaley-Cattell (1956) study being a notable exception.

Until recently, of course—i.e., before electronic computers became readily available—the computation of factor scores was a task of such Herculean proportions that an investigator could hardly be blamed for not planning his study to include these procedures. But this is no longer true. Computers reduce the task to easily managed proportions.

Present indications are that factor analysis will continue to be a popular research tool. It may very well become more popular. Psychometricians working with the method have known for some time that as a means for reducing the number of variables, for defining more meaningful or fundamental dimensions, for increasing the internal consistency of principal variables and for other reasons, factor analysis can be quite useful, particularly in the initial stages of investigations which use analysis of variance and similar methods in later stages to study the effects of many potential influences on many dependent variables. More and more social scientists are be-

coming aware of these kinds of applications of factor analytic procedures. Many of these applications involve computing factor scores.

It would seem, therefore, that some of the formerly academic questions about the characteristics and relative merits of the various procedures for estimating factor scores are now, or will soon become, practical research questions of vital importance. "Which method is best for a particular purpose?" "Does it make much difference?" "Are the differences in results in two studies of the effects of A, B, and C on D interpretable in terms of the different methods used to calculate factor scores in the two studies?" The present article is an exploratory study designed to provide some preliminary answers to questions like these.

Methods to be Considered

Basis for Classification of Methods

The available techniques for estimating factor scores can be roughly classified as relatively more *complete* or less complete—i.e., *incomplete*. The former are also frequently referred to as "exact" methods, whereas the later are termed "crude," "approximate," or "inexact."

Briefly stated, the distinction drawn has to do mainly with whether or not a least-squares procedure is employed and the amount of information in the factor coefficient matrix (i.e., the structure or pattern) which is used. What are referred to as complete methods use the factor coefficient matrix, as it is given by the usual calculations in factor analysis, and also employ a least-squares procedure, involving computation of an inverse, to minimize error of estimation; they differ in the way they define this error of estimation. By contrast, the incomplete methods do not use a least-squares procedure and may not use all of the information provided by the factor coefficient matrix.¹

¹ It may be noted, parenthetically, that by virtue of not using the least-squares procedure the so-called "crude" methods are less susceptible to shrinkage effects on the reliability of measurement in cross-validation and hence are, in this sense, the more "exact" methods.

Complete Methods

The best known of the complete methods states the problem as one of linear multiple regression analysis in which the common-factor score for person i on factor j is to be estimated by

$$f_{ji} = B_{j1}Z_{i1} + B_{j2}Z_{i2} + \cdots + B_{jn}Z_{in}; \quad (j = 1, 2, \dots, m) \quad (1)$$

where the n terms, Z_{ik} ($k = 1, 2, \dots, n$), symbolize the obtained scores on n tests, expressed in standard score form, and the B_{jk} are beta weights to be determined by minimizing the sum of squares of the discrepancies between the hypothetical true factor, f'_{ij} , and the estimate, f_{ij} . The normal equations (involving the intercorrelations of the n variables and the correlations between factors and variables) then follow and the solution turns out to be merely a special case of the most commonly met least-square multiple regression equation. In matrix form this is, for the orthogonal case,²

$$F_1 = ZR^{-1}A = ZE_1 \quad (2)$$

where, for N subjects, n original variables and m factors, F is an N by m matrix of factor scores computed by this, the first-mentioned method (hence the subscript 1), Z is an N by n matrix of scores on the original variables in standard score form, R^{-1} is an n by n inverse of the matrix of intercorrelations among variables, A is an n by m matrix of factor coefficients.

In equation (2), and in other factor score estimation procedures to be described, all to the right of the standard score matrix is summarized in a single n by m matrix of weights, E , termed the factor estimation matrix. Cattell (1962) has suggested that this matrix, rather than the A matrix which is characteristically used, might furnish the basis for the interpretation of factors. As will be seen at a later point in this paper, the E matrices for the various methods of computing factor scores are also the crucial quantities needed for comparisons of the different methods.

Bartlett's (1932) method is not as well known as the above. It makes the sum (over n variables) of the squares of the unique fac-

² Discussion and analysis throughout this study refer only to orthogonal factors solutions.

tors a minimum. The derivation begins with what is often termed the "fundamental equation of factor analysis," viz.,

$$Z_{ki} = a_{k1}f_{1i} + \cdots + a_{km}f_{mi} + U_k u_{ki}; \quad (k = 1, 2, \dots, n) \quad (3)$$

solves for the unique factor score, u_{ki} , and minimizes the sum of squares of these, yielding the formula

$$F_2 = ZU^{-2}A(A'U^{-2}A)^{-1} = ZE_2, \quad (4)$$

where U without subscript is the diagonal matrix of uniqueness coefficients and the other symbols are as defined previously.

Those who are familiar with factor score estimation procedures will recognize that in the case of orthogonal factors the estimation achieved by (4) differs from the shortened method developed by Ledermann (1939) only in the last term on the right: in Ledermann's procedure this is $(I - A'U^{-2}A)^{-1}$. A correlational comparison of method F_1 with Bartlett's method is thus essentially also a comparison of F_1 with Ledermann's shortened method.

A third "complete" procedure, sometimes labeled the "direct" method or the method of ideal variables (Harman, 1960, p. 360), also begins with the fundamental equation (3) above. However in this derivation it is assumed that the unique part of the score, Z_{ki} , can be ignored or that a good approximation to the obtained score is given by consideration of only the common-factor portion. With this assumption, the fundamental equation can be written in matrix form as

$$Z = FA' \quad (5)$$

whence, by multiplying on the right first by A , then by $(A'A)^{-1}$, still another least-squares estimate of factor scores is obtained (Horn, 1964), as given by

$$F_3 = ZA(A'A)^{-1} = ZE_3. \quad (6)$$

Analytically, then, the three so-called complete methods considered here differ primarily in that (1) the first uses the matrix of correlations between variables to get the regression of each factor on the n variables, in contrast to the other two methods, (2) the second minimizes the unique factors in contrast to the other two, and (3) the third assumes that the obtained scores contain *only* common-factor components, no unique components. Thus, from

a consideration of analytic properties alone, it might be expected that, particularly when unique components are substantial, the three methods can give corresponding arrays of factor scores which correlate rather lowly.

Incomplete Methods

The so-called incomplete methods, developed for use in the days before electronic computers became generally available, were designed to eliminate computations like those involved in obtaining an inverse. One of the most commonly used of these techniques merely treats the factor coefficients (i.e., the factor "loadings") as if they were the weights to use in estimating factors from variables (rather than, as in fact they are in the factorial model, the weights to use to estimate variables from factors). Under this assumption the computing formula can be represented by

$$F_s = ZA = ZE_s, \quad (7)$$

which avoids inversion entirely but does use the intact factor coefficient matrix, A .

An even simpler procedure is to give nonzero weights only to those variables which are salient (by some arbitrary criterion) in the factor, as revealed by the factor coefficient matrix. The computing formula for this (the F_s) estimate is like that given in (7) except that all nonsalient factor coefficients in A are replaced by zeros; the other weights in A are not altered.

The method that is most frequently used is merely to sum the standard scores on the variables that are salient in a given factor. This is represented in a matrix equation by setting all nonsalient coefficients in A at zero, all salients at 1.0 and carrying out a matrix multiplication like that indicated in (7). The matrix of scores obtained by this procedure is symbolized F_s in subsequent tables.

Procedures and Results

The similarity between the different methods for estimating factor scores might be described either by use of what is sometimes called the "coefficient of congruence" (cf. Burt, 1941; Tucker, 1951) or by intercorrelating corresponding factor score estimates. The first procedure uses the E_i matrices alone and is merely the inner product of two columns of the E matrices adjusted by the

product of the "lengths" of the two vectors implied, i.e., the full matrix of coefficients of congruence for estimates E_1 and E_2 is given by

$$C = (E'_1 E_1)^{-1/2} (E'_1 E_2) (E'_2 E_2)^{-1/2}. \quad (8)$$

If E_1 and E_2 are arranged so that weights for estimating a given factor appear in the same column in both matrices, the main diagonal of C will contain coefficients of congruence for estimates of the same factor.

Now in fact the coefficients of congruence are equivalent to the correlations between the factor score estimates and thus provide merely a shortened method for computing these latter. Because the simple term "correlation" is more widely understood, this word rather than "coefficient of congruence" is used in subsequent discussion.

Two samples of subjects, variables, and factors were used, although not all methods of factor score estimation were tried out in both sets of data. Some of the same procedures were employed in both analysis, however, thus permitting comparisons across samples.

In the first analysis factoring proceeded with 59 ability, motivational, and general personality measurements obtained on 297 male and female adults; 16 principal axes factors were computed and then rotated to a position given by application of the varimax criterion. The matrix obtained from this rotation was used as the basic A matrix in calculations of the factor score estimates symbolized above by F_1, F_2, F_3 , and F_4 . In the second analysis an entirely different sample of 137 male and female adults was drawn; 52 personality measures (most of which were different from those used in the first analysis) were obtained on each subject; 18 centroid factors were extracted; the varimax criterion was again applied, and the factor score estimates symbolized by F_1, F_4, F_5 , and F_6 were computed.

The similarity between methods is indicated by the coefficients given in Table 1. The correlations between corresponding factor score estimates are shown in the upper sections of the table. The simple average correlation in a column is shown in the first row at the foot of the column. Just below these values are shown the simple averages of the absolute values of the correlations which a

TABLE 1
Correlations between Corresponding Factor Score Estimates Based upon Different Procedures

Factor Number	Analysis 1		Analysis 2		Analysis 1		Analysis 2		Analysis 2	
	F ₁	F ₂	F ₃	F ₄	F ₁	F ₂	F ₃	F ₄	F ₅	F ₆
1	96	98	82	80	78	76	87	84	93	98
2	98	97	87	86	81	78	98	88	96	98
3	82	78	66	86	84	77	98	78	93	97
4	96	97	84	81	76	78	98	86	90	89
5	98	95	80	93	87	80	97	82	93	96
6	98	97	82	95	86	87	98	84	91	98
7	98	98	84	93	85	82	99	87	83	97
8	96	98	80	87	84	83	98	82	93	98
9	95	96	78	88	89	89	99	85	99	99
10	98	94	76	76	76	73	91	87	98	99
11	98	96	83	91	81	78	98	85	89	99
12	95	97	79	87	80	78	97	84	92	98
13	95	97	82	90	90	88	94	85	98	99
14	93	95	82	93	79	76	92	86	88	98
15	93	97	79	91	83	80	95	83	90	98
16	89	89	73	92	75	70	97	79	81	96
17				91	76	73			89	98
18				85	84	81			96	98
Average Correspond.	94	95	80	88	82	79	97	83	92	97
Average Non-Correspond.	08	09	07	14	15	13	15			

given factor estimate had with all other factors estimated by the same procedure. Since the factor solutions were orthogonal, these averages should be near zero. The extent to which they depart from zero indicates the extent to which the factor estimation procedure departs from the assumptions of the factorial model.

Discussion

Choice between the methods in any particular study must rest upon a number of considerations. Some of these will be purely practical. For example, one would normally need to consider the size of the computer that is available, the programs already developed for this computer, the number of variables, the number of subjects, etc.

On the 7090, where the above work was done, there is virtually no difference between the last three methods in terms of time on the computer. But methods F_5 and F_6 involved more clerical work because the computer was not used to modify the A matrix before going to the multiplication of the form of (7). Of course, programs could be developed to perform the clerical chores involved here and, once programmed, the operations would take very little time on the computer.

The most time-consuming (on the machine) method turned out to be F_1 followed closely by F_2 , whereas F_3 proved to be only slightly slower than any one of the incomplete methods. This difference in time refers only to that involved in obtaining the E matrices; the multiplication from this point on is as fast for the complete methods as for the incomplete methods. When a large number of subjects is involved the time required to obtain an E matrix is only a very small fraction of the total time required to get the factor scores.

On purely theoretical grounds one can argue for virtually any one of the methods. The F_1 method gives a good least-squares estimate of the common-factor, for example; yet there is something to be said for a method that minimizes that error of measurement in the unique component, as implied by the F_2 method; and the F_3 method is more nearly the direct solution for true factor scores. The computations here might also be applied with the U matrix, thus giving the complete set of factor scores. But psychometrists have long known that least-squares procedures, such as those of the

three complete methods, imply shrinkage of reliability in cross-validation, particularly so when the number of variables is large relative to the number of subjects and when the variables are rather highly intercorrelated, as is often the case in factor analytic studies. Hence on theoretical grounds one might prefer one of the incomplete methods, particularly F_2 because it does least to take advantage of the chance influences affecting the weights in the A and E matrices. The fact that the correlations between factors are not accurately retained by this method would need to be given serious consideration, of course. By prudent choice of suppression variables these correlations can be closely approximated, however, although much trial-and-error is sometimes necessary before satisfactory suppressors can be found and crossvalidation becomes about as crucial when these procedures are resorted to as when the least-squares procedures are used.

Thus many pro and con arguments can be presented for any one of the methods. The purpose here is not to argue the advantages of any particular method but merely to call attention to some of the various possibilities which do exist and to present descriptive data for use in comparing the methods. Depending upon his practical circumstances and his theoretical inclinations as dictated by the nature of his data, an investigator may select the method which is most appropriate for his particular study. If one feels that the F_1 method is best on theoretical grounds, for example, but the inversion program available on his computer will not handle a matrix of the order of his R , or is too slow with this size of matrix, the F_2 method might be selected as giving an adequate approximation, involving inversion of only a small matrix. On the other hand if sample size were small relative to the number of variables, and only a very few variables were salient in some factors, the method F_3 might be selected as most appropriate.

Summary

Six methods for computing factor scores were described and some of the characteristics of each method were discussed. Using two samples of subjects, variables, and factors, the methods were used to obtain score estimates and these were intercorrelated. The correlations between corresponding estimates were found to range from .70 to .99; the simple average correlation between different

methods ranged from .79 to .97. The methods which used the factor coefficient matrix and a least-squares estimation procedure were intercorrelated above .90. The methods which did not use a least-squares procedure likewise correlated above .90 with each other. The correlations between these two broad classes of methods averaged between .79 and .88. The methods were roughly compared in terms of the time and programming demands they made, and it was suggested that these practical matters might still need to be given consideration, along with theoretical matters, in the actual choice of a factor estimation method for a particular study.

REFERENCES

- Baggaley, A. and Cattell, R. B. "A Comparison of Exact and Approximate Linear Function Estimates of Oblique Factor Scores." *British Journal of Statistical Psychology*, IX (1956), 83-86.
- Burt, C. *The Factors of the Mind*. New York: Macmillan, 1941.
- Bartlett, M. S. "The Statistical Conception of Mental Factors." *British Journal of Psychology*, XXVIII (1937), 97-104.
- Cattell, R. B. "The Basis of Recognition and Interpretation of Factors." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXII (1962), 667-697.
- Harman, H. H. "On the Rectilinear Prediction of Oblique Factors." *Psychometrika*, VI (1941), 29-35.
- Harman, H. H. *Modern Factor Analysis*. Chicago: University of Chicago Press, 1960.
- Horn, J. L. "A Note on the Estimation of Factor Scores." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIV (1964), 525-527.
- Horst, P. "The Prediction of Personal Adjustment." Social Science Research Council Bulletin, No. 48, 1941.
- Kestelman, H. "The Fundamental Equation of Factor Analysis." *British Journal of Psychology, Statistical Section*, V (1952), 1-6.
- Kaiser, H. F. "Formulas for Component Scores." *Psychometrika*, XXVII (1962), 83-85.
- Ledermann, W. "On a Shortened Method of Estimation of Mental Factors by Regression." *Psychometrika*, IV (1939), 109-116.
- Thomson, G. H. "On Estimating Oblique Factors." *British Journal of Psychology, Statistical Section*, II (1949), 1-2.
- Thurstone, L. L. *Multiple-Factor Analysis*. Chicago: University of Chicago Press, 1947.
- Tucker, L. R. "A Method for Synthesis of Factor Analysis Studies." Department of the Army, Personnel Research Section Report, No. 984, 1951.

TIME-SAVING PROCEDURE FOR COMPUTING Z SCORES¹

HELEN A. HEATH

Institute for Psychosomatic and Psychiatric Research and Training,
Michael Reese Hospital

THIS paper describes a simplified procedure for transforming a distribution of raw scores into standard (Z) scores which have a mean of 50 and a sigma of 10. Once the standard score for the first subject has been obtained, the others may be determined rapidly with each Z value being utilized in the computation of the subsequent one.

Since the shapes of the raw score and standard score distributions are identical, increases or decreases between adjacent values are proportional. For example, if B has a reading three points higher than A, and C's reading is six points higher than B's, the difference between Z_B and Z_C will, likewise, be twice as great as the difference between Z_A and Z_B . The following derivation reveals that numerical difference between adjacent standard scores is equal to $10/\sigma(X_2 - X_1)$.

By definition,

$$Z_1 = \frac{10(X_1 - \bar{X})}{\sigma} + 50 \quad (1)$$

and

$$Z_2 = \frac{10(X_2 - \bar{X})}{\sigma} + 50. \quad (2)$$

¹This study was supported by the State of Illinois Mental Health Fund, Grant 1711.

If (1) is subtracted from (2), we have

$$Z_2 - Z_1 = \frac{10(X_2 - \bar{X})}{\sigma} + 50 - \frac{10(X_1 + \bar{X})}{\sigma} - 50. \quad (3)$$

By cancelling the \bar{X} 's and 50's, simplifying and transposing Z_1 , the equation becomes

$$Z_2 = Z_1 + 10/\sigma(X_2 - X_1). \quad (4)$$

Illustration of Method

Table 1 consists of a worksheet designed to obtain Z scores by means of equation (4). Before proceeding, it is necessary that the mean, standard deviation, $10/\sigma$ for the distribution, and the Z score for the first subject be available. Equation (1) may be used for computing this Z score.

TABLE 1

Work Sheet

Column 1 Subjects	Column 2 Raw Scores	Column 3 Difference Scores	Column 4 Z Scores
1	30		51.66283
2	26	-4	37.045
3	32	6	58.972
4	34	2	66.281
5	28	-6	44.354
6	29	1	48.008
7	29	0	48.008
8	27	-2	40.699
9	31	4	55.317
10	26	-5	37.045
11	33	7	62.627
	Mean	29.545	
	σ	2.7363	
	$10/\sigma$	3.65457	

The difference between each value and the preceding one in column 2 is recorded in column 3. Since 26 is 4 points less than 30, the entry which refers to the second subject is -4. At this time, column 4 will contain only Z_1 , 51.66283. The procedure as outlined below assumes the use of a Marchant Calculator with two keyboards; however, it may be adapted to other types of desk computers.

1. Have carriage as far to your left as possible.
2. Set the machine decimal an equal number of spaces from the right on the keyboard and on the dial which designates answers

to multiplication and addition problems. For this illustration, we are using five places.

3. Punch in the first Z score and press the add bar.
4. Lock the value of $10/\sigma$ in the keyboard using the same decimal setting.
5. Multiply by the first difference score which in this illustration is -4 . The dial previously reading 51.66283 will now read 37.04455 which is Z_2 . Record in column 4, rounding as desired, and shift carriage back to extreme left.
6. Retain the values in the dial and in the keyboard, and continue multiplying by the difference scores from column 3. Each multiplication will yield a Z score. Be sure that the carriage is returned after every multiplication.

A slight adjustment is required when the raw scores contain decimal values. For each number to the right of the decimal, the $10/\sigma$ should be located one additional space to the right on the keyboard. To illustrate this we may assume that a decimal point was placed in the center of the two digit raw scores on Table 1. (That is, 30 would become 3.0, etc.) The Z scores would remain the same; however, the $10/\sigma$, carried to five places, would be 36.54570 and each difference score would be a decimal value. It is evident that in order to continue obtaining the same Z scores, the number of decimal places in the multiplicand would need to be reduced to compensate for the one place in the multiplier. Consequently, the decimal point in the keyboard should be four places from the right instead of five. That is 36.5457 should be punched in instead of 36.54570.

Results may be rapidly checked by computing the mean of the column of standard scores. If correct, this will approximate 50.000. The reader should be cautioned that this check detects errors only in the computations directly involved in obtaining Z scores. The σ and $10/\sigma$ must be checked independently.

RELIABILITY AND VALIDITY: BASIC ASSUMPTIONS AND EXPERIMENTAL DESIGNS

EDWARD E. CURETON
University of Tennessee

IN the derivations of all formulas for test reliability and validity, the basic assumptions are that, in the population, the errors of measurement in all forms of all tests and criterion measures are statistically independent both of one another and of the true scores. If we confine attention to the linear case, we can read "uncorrelated" for "statistically independent." Sampling of the literature indicates that few investigators realize that these assumptions imply definite restrictions upon the experimental designs which are acceptable in research based on psychological and educational test scores when the test reliabilities are to be taken into account.

Consistency, Stability, and Inter-Form Reliability

There are two main types of errors of measurement which are important in objective testing. The first type, which has received exhaustive treatment in the literature, may be termed *inconsistency*. It results from the fact that every form of a test contains a finite number of items. Hence, one examinee will just happen to be able to produce the right answers to more items of Form A, and another to the items of Form B, even though both forms are equally difficult on the average in the population of which these examinees are members.

The second type of error, which is hardly less important than the first, may be termed *instability*. It results from the fact that an examinee will not give the same set of responses to the same set of items at two different times, however small the time difference may be. When two forms are administered, the decrement in inter-form cor-

relation is continuous over time, including the time necessary to react to the items of one form, and is numerically quite substantial. In the experimental literature this type of unreliability is very largely neglected, and the writer has never seen a stability coefficient in a printed test manual. Such being the case, the statements above that instability is hardly less important than inconsistency, and that decrement in inter-form correlation is substantial even over short time intervals, appear to require some documentation.

Twelve 250-item forms of a number-checking test were administered in consecutive five-minute intervals to 167 tabulating equipment operators (Taylor, Manson, and Stone, 1945), and all inter-form correlations were computed. We define the *separation interval* between two forms as 1 if they were administered consecutively, as 2 if one other form intervened, . . . , as 11 when referring to the single correlation between forms 1 and 12. For separation interval 1 there are eleven correlations ($r_{12}, r_{23}, r_{34}, \dots, r_{11-12}$); for separation interval 2 there are 10 ($r_{13}, r_{24}, r_{35}, \dots, r_{10-12}$), etc. For the 11 different separation intervals, the average correlations are given below:

Sep. Int:	1	2	3	4	5	6	7	8	9	10	11
No. r 's:	11	10	9	8	7	6	5	4	3	2	1
Av. r :	.925	.901	.880	.858	.839	.818	.782	.758	.706	.649	.583

The numerical value of the drop, from .925 to .583, is larger here than it would be for unspeeded tests because of the greater differential boredom and fatigue effects, but the important point is that *every* average r is larger than the one following, and the decrements form a remarkably smooth sequence.

The same test was administered also, as a part of the same study, to 107 stencil typists and key-punch operators. For this group the average correlations were:

Sep. Int.	1	2	3	4	5	6	7	8	9	10	11
No. r 's:	11	10	9	8	7	6	5	4	3	2	1
Av. r :	.915	.905	.886	.861	.836	.813	.788	.765	.723	.685	.628

Here again every average r is larger than the one following, and again the decrements form a quite smooth sequence.

Three 27-item forms of the A. C. E. Opposites Test were mimeographed on one sheet and administered to 187 undergraduate students in five classes on a Friday. Three other forms, mimeographed also on one sheet, were given to some of the same classes on the fol-

lowing Monday, some on the following Wednesday, and some on the following Friday (Cureton, 1939). In each case the tests were given without time limit. The average of the six correlations between forms administered at the same period was .729. The average of the nine correlations between forms administered on two different occasions (averaging about five days apart, including a weekend) was .517. Note that this is the mean reliability of *one* 27-item form.

The stability coefficient, over a given time interval, is the inter-form correlation over the same interval divided by the consistency coefficient (Cureton, 1958). The value .729 is somewhat lower than the actual consistency because the three forms were administered serially at each sitting rather than simultaneously. If we divide .517 by .729 we obtain the figure .709 as an upper limit for the stability coefficient. If the consistency were as high as .800, the stability would be as low as .646. The one-week stability coefficient is lower than the consistency coefficient, but not much lower.

The term "reliability" is a generic term. When a numerical coefficient is reported, it should be termed a consistency coefficient, a stability coefficient, or an inter-form reliability coefficient. When consistency coefficients are termed reliability coefficients without further qualification, as is now quite generally the case, naive and even not-so-naive test users tend to assume that the Kuder-Richardson and split-half, Spearman-Brown methods are merely experimental and computational alternatives to the inter-form correlation method. Explicit statements of the method of computation used, and even general cautions that the former methods yield numerically higher coefficients, do not dispel the confusion. It is too late now to reserve the term "reliability coefficient" for the inter-form correlation, but it should be emphasized in every test manual that unreliability includes instability as well as inconsistency, and that *only* the inter-form correlation is reduced by both.

The Estimation of Consistency

When a consistency coefficient is computed from the split-half correlation and the Spearman-Brown formula, the two half-tests must be administered *simultaneously*. The time displacement which results from correlating odds against evens is not ordinarily serious when the number of items attempted is even, and it is zero when

this number is odd. Cronbach (1951) has shown that α (KR-20) is exactly equivalent to the split-half, Spearman-Brown consistency coefficient when the split is perfect: i.e., when it is such that the two half-tests are equally reliable and the errors of measurement do actually correlate zero with each other and with the true scores.

In some types of tests the items come in groups: e.g., paragraph-reading tests with several items on each paragraph, and table-reading tests with several items on each table. Differences among examinees in general comprehension of a paragraph or table then produce correlated errors of measurement (of the inconsistency type) between items based on the same paragraph or table. Each paragraph or table, with all items based on it, must therefore be assigned to one half-test to avoid correlated errors *between* half-tests; correlated errors *within* half-tests do not violate the basic assumptions. In this situation an odd-even split of the *superitems*, each consisting of a paragraph or table and all associated items, may itself be inadequate if the number of such superitems is even. If this number is a multiple of 4, patterns such as the following may be used:

A—superitems:	1,	4,	5,	8
B—superitems:	2,	3,	6,	7,

Inequalities in time-variance are much less important here than inequalities in mean time of administration.

When the number of superitems is odd, the two half-tests will be of unequal length if the number of items is the same for each superitem. If item-analysis data are available, the resulting inequalities in consistency can be reduced to some extent by assigning to the shorter half-test those superitems whose mean item-discriminations are highest. This procedure, which is a matter of the layout of the test booklet, should be noted by test constructors. It is *essential* when the number of superitems is 3, if the consistency of the test is to be computable. The Spearman-Brown formula is derived on the assumptions that the half-tests are equally consistent and equally variable, and while these requirements are not stringent (Cureton, 1958), they should not be violated to the extent implied by two half-tests which are equally consistent and equally variable *per item*, but one of which is twice as long as the other.

If there are more than three superitems, the generalized α (KR-

20) may be used instead of the split-half, Spearman-Brown procedure in estimating consistency. The formula is

$$\alpha = \frac{k}{k-1} \left[1 - \left(\sum_1^k \sigma_i^2 \right) / \sigma_t^2 \right],$$

where k is the number of superitems, σ_i^2 is the score-variance on each superitem (no longer 0 or 1), and σ_t^2 is the variance of the scores on all superitems (i.e., the variance of the total scores).

To find the consistency of a speed test, we face the logical dilemma that the half-tests or forms should be administered simultaneously but with experimentally independent time limits. We can avoid this dilemma if there are at least *five* equivalent forms, and these forms are administered serially at equal time intervals (preferably consecutively at one sitting). Equivalence here implies fairly closely equal consistencies and variances, but not equal mean difficulties, and the experimental design requires separate and equal time limits for all forms. In this case we can compute the average correlations for at least four separation intervals from the inter-form correlations. We then plot average correlation against separation interval, draw a smooth curve as nearly as possible through all the plotted points (with particular attention to the first four if there are more than four), and extrapolate backward one unit to separation-interval 0. The corresponding value on the average- r scale is then the consistency coefficient of the speed test. Four separation intervals (hence 5 forms), provide one degree of freedom for fitting a second-degree curve; if there are only three points there is no way to estimate the extent of the irregularities in the average correlations and hence of the limits of variability in fitting and extrapolation.

This technique was devised originally by Feldstein (Newstetter, Feldstein, and Newcomb, 1938). For the number-checking test data on 167 tabulating equipment operators, the result is a consistency coefficient of .94, and no reasonable variation of the visual extrapolation can change this value by as much as .01. For the data on 107 stencil typists and key-punch operators the result is a consistency coefficient of .93, and again no reasonable variation of the visual extrapolation can change this value by as much as .01. It should be noted that in the absence of a rational equation for the separation interval-average r function, visual fitting (using, e.g., a ship curve) gives better results than does fitting any of the simpler

empirical equations. When a parabola was fitted by least squares to these data, the fit was obviously much worse, it was improved only slightly by weighting each average r by the number of forms on which it was based, and it was about equally good (but no better than visual fitting) when each average r was weighted by the number of correlations averaged.

The Estimation of Stability and Inter-Form Reliability

Since the decrement in inter-form correlation is continuous over time from the instant the examinees start to work on the first form administered, *there is no single population value for either the stability coefficient or the inter-form reliability coefficient.* The population values estimated by inter-form correlations and inter-form correlations divided by consistency coefficients are themselves functions not only of the length of the time interval separating the first administration from the second, but also of the *particular* time interval. Since it is highly probable that there are differential diurnal variations in the reactions of examinees to test items, the particular times of day at which the two forms were administered has some effect, and particularly the matter of whether the two forms were or were not administered at the same time of day. (By "differential," we mean different for different examinees. The average elements in changes, which raise or lower all raw scores equally, have no effect on correlations.) It is also highly probable that the particular events which occur during the time interval may have some effect, so that a 4-day interval from Monday to Friday may produce more or less instability than a 4-day interval including a weekend. It would seem to follow, then, that when an author reports an inter-form reliability coefficient or a stability coefficient, he should state not only the length of the time interval, but also ideally the time of day and the date (including the day of the week) on which each form of the test was administered, and include in addition in some cases a description of such intervening events (e.g., final examinations, epidemics, snowstorms, etc.) as might be expected to produce unusual differential changes in stability.

Over a given time interval, the correlation between two forms of a test is the geometric mean of their inter-form reliabilities. Only when the two forms are equally reliable is it the inter-form reliability of each form separately.

The stability coefficient is a special case of the correlation corrected for attenuation. If r_{11} is the inter-form correlation between two parallel forms of the same test (not necessarily equally consistent or equally variable or equally difficult), c_1 and c_I are their consistency coefficients, and s_{11} is the stability coefficient, we have,

$$s_{11} = r_{11} / \sqrt{c_1 c_I}.$$

If, in addition, the two forms are equally consistent, $c_1 = c_I = c_{11}$, say, and

$$s_{11} = r_{11} / c_{11}.$$

The stability coefficient, over a given time interval, is the value which the inter-form reliability coefficient, over the same time interval, would take if both forms were perfectly consistent. The two forms, moreover, will be equally reliable over *any* time interval if and only if they are equally consistent.

Since the stability coefficient is the inter-form reliability coefficient corrected for inconsistency, its numerical value is independent of the length of the test, i.e.,

$$s_{11} = \frac{r}{c} = \frac{R}{C},$$

where r and c refer to short forms of the test, and R and C refer to forms n times as long. The single c and C indicate that we have assumed that the forms of each pair are equally consistent. By the Spearman-Brown formula,

$$C = \frac{nc}{1 + (n-1)c}.$$

We then have

$$\frac{r}{c} = \frac{\frac{R}{nc}}{\frac{1 + (n-1)c}{1 + (n-1)c}} = \frac{R[1 + (n-1)c]}{nc}.$$

Solving for R

$$R = \frac{nr}{1 + (n-1)c}.$$

The Spearman-Brown formula does not apply to the inter-form reliability coefficient. It applies only to the consistency coefficient.

When both forms are lengthened (or shortened) by a factor n , the estimate of the inter-form reliability of the lengthened (or shortened) test is given by the last formula above. Failure of measurement workers (including the present writer!) to concern themselves explicitly with the differences between consistency, stability, and inter-form reliability has permitted this point to escape notice for more than 50 years.

When the same test (other than, perhaps, a pure speed test) is administered a second time, the test-retest correlation is neither a stability coefficient, a consistency coefficient, nor an inter-form reliability coefficient. It is not a consistency coefficient because the item set is the same at both administrations and the time interval is not zero. It is not an inter-form reliability coefficient because the item set is the same on both occasions. It is not a stability coefficient first because the item set is finite (tending to make its value lower than that of the stability coefficient), and second because perseveration effects (including memory on the second occasion of how some of the items were marked on the first occasion) introduce correlated errors of measurement that tend to raise its value. So far as the writer can determine, the test-retest correlation has *no* clear interpretation.

Loveland (1952) administered Form A of seven of the Differential Aptitude Tests (omitting the Clerical Speed and Accuracy Test) without time limits to all available students in a four-year high school in Knox County, Tennessee. The testing was done in two sessions on two consecutive mornings. Eight days later this testing program was repeated, using the same forms of the tests. Neither teachers nor students knew in advance that the re-test session was to be held. Due to absences, including a student excursion to Washington, D. C., the number of students varied from test to test. The odd and even items of each answer sheet were scored separately. The results are given in Table 1. In this table, the subscripts 1 and I refer to the first session; 2 and II to the second. The Arabic numerals 1 and 2 refer to scores on the odd items; Roman numerals I and II to scores on the even items. In the rows at the bottom, the c 's are consistency coefficients, the r 's are inter-form reliability coefficients, the s 's are stability coefficients, and the tr 's are test-retest coefficients.

In every column except the last, the stability coefficient is highest, the consistency coefficient is second, the test-retest coefficient is third,

TABLE 1

*Consistency, Stability, Inter-Form Reliability, and Test-Retest Correlations
for Seven of the Differential Aptitude Tests*

	Verbal Reas.	Abstr. Reas.	Spell- ing	English Usage	Numer- ical	Space	Mechan. Reas.
r_{1I}	.8308	.8319	.8770	.8260	.7580	.8309	.7558
r_{2II}	.8677	.8568	.8746	.8293	.7994	.8673	.7706
r_{12}	.8355	.7721	.8857	.7954	.7939	.8332	.8224
$r_{I II}$.8556	.7808	.8628	.7735	.7339	.8288	.8196
r_{12}	.7952	.7675	.8160	.7574	.7067	.7979	.7394
$r_{1 II}$.8155	.7267	.8478	.7181	.7115	.7767	.7294
N	572	583	417	605	595	605	587
$c = \sqrt{r_{1I}r_{2II}}$.849	.844	.876	.828	.778	.849	.763
$r = (r_{12} + r_{I II})/2$.805	.747	.832	.738	.709	.787	.734
$s = r/c$.949	.845	.950	.891	.911	.927	.962
$tr = (r_{12} + r_{I II})/2$.846	.776	.874	.784	.764	.831	.821

and the inter-form reliability coefficient is lowest. For Mechanical Reasoning, however (last column), where a high degree of perseveration might well be expected, the test-retest coefficient is higher than the consistency coefficient. Perseveration effects, or effects of some related type, can also be seen in the two rows at the top: for every test except Spelling, the consistency coefficient is higher at the second administration than at the first. Note also that in these data the stability coefficients are higher than the consistency coefficients, while in the case of the A. C. E. Opposites Test, the stability coefficients were lower than the consistency coefficients. This may be due to the differences in the experimental designs: in Loveland's study each inter-form correlation was the correlation between one half-test given at the first administration and the other half-test given for the second time at the second administration. This is not the best experimental design for the determination of stability coefficients; differential practice effects, perseveration effects, and whatever caused the consistency coefficients to be generally higher at the second administration than at the first may have inflated the stability coefficients computed from these data.

The Estimation of Criterion Validity

There are two fundamentally different types of criteria. The first may be termed *sui-generis* criteria. They exist, quite apart from any efforts to predict them, they are worth predicting, and their reliabil-

ities may be high or low. Examples include persistence in college, success or failure in a training course, dollar volume of sales, years of service in a company, etc. The unreliability of the criterion measure or attribute sets a natural upper limit for the validity of any possible predictor or combination of predictors. The validity of any predictor or predictor battery is its raw correlation or multiple correlation with the criterion. We may call such a correlation an index of *raw validity*.

The second type of criterion may be termed a *constructed* criterion. We start with a trait-name: e.g., academic ability (at such-and-such educational level), job proficiency, or sales accomplishment. A corresponding criterion measure is then constructed. For academic ability it might be a grade-point average based on academic courses only. For job proficiency it might be an index based on units of work completed or items produced, corrected by some more or less arbitrary formula for number of errors or amount of material spoilage, or it might be a rating by the supervisor or an average of two or more such ratings. For sales accomplishment it might be an index based on dollar volume of sales and number of new customers added, with perhaps a correction for the difficulty of the territory.

In any event, the constructed criterion must be acceptable as an *operational definition* of the trait-name, and if it is factorially complex, the factors entering into it must be assumed to correlate sufficiently with one another to make the single numerical index meaningful and useful.

Since the error of measurement of a constructed criterion measure is no part of an operational definition of any substantive trait, it is evident that when we use a constructed criterion we are attempting to predict the *true* criterion scores, not the *raw* scores. If x represents a single predictor score or a composite of several predictor scores weighted by regression coefficients, and y_o and y_o represents scores on two parallel forms of a constructed criterion measure, with true score y , the estimate of the correlation or multiple correlation between the predictor variable or battery and the true criterion is

$$r_{xy} = \sqrt{r_{xo}r_{xo}/r_{o0}}.$$

Since we are predicting true criterion scores, such a correlation might well be termed an index of *true validity*.

The two forms of the criterion measure must be experimentally

independent, they must measure the same trait or combination of traits, and they must cover the same time period of criterion behavior. On the other hand, they need not be equally consistent, equally variable, nor expressed in equal units. This is fortunate, for in practice it is often difficult enough to obtain two parallel criterion measures which cover the same period of criterion behavior and are still experimentally independent.

We can recognize two types of true validity, which may be termed concurrent true validity and forecast true validity.

In the case of concurrent true validity, the criterion measure is usually one which is very hard to obtain, and we want to know how valid the test or battery is as a *substitute measure* of the criterion variable. In this case the test or battery must be administered at the *center* of the time interval over which the two sets of criterion behavior are observed or measured, so that both the test or battery and the two forms of the criterion measure are essentially (though not entirely) free from instability errors. This is the only design satisfying the requirement that the time interval separating the administration of each pair must be the same. In this case they will be the same *on the average*, but within-period instability will still produce some additional attenuation of r_{oo} .

In the case of forecast true validity, the predictor test or battery should be administered at the administratively natural time (e.g., at or shortly before entrance to college, admission to a training course, or employment), and the criterion data should cover the period over which they will be most valid in terms of the criterion trait concept. The two forms of the criterion measure must still cover the same time period, but the instability error resulting from the earlier administration of the predictor test or battery is intrinsic to the prediction enterprise.

When we use the so-called "present-employees" method of validating a predictor test or battery we are not logically carrying out a concurrent validation study, even though the time of administration of the predictor may coincide more or less closely with the criterion observation period. The objective is still a forecasting objective rather than a criterion-measure substitution objective, but for practical reasons (usually to get the job done in a short time) the investigator adopts a less-than-optimum design and hopes that the spurious inflation of the resulting index of true validity (due to

reduction of the instability error) will not be serious, and that the effects of the criterion experience upon the predictor scores will not be serious either.

Correction for Attenuation

When a correlation is corrected for attenuation, the objective is ordinarily that of estimating an intrinsic relationship: the correlation between the true scores on two tests or between the true scores on a test and a criterion measure. In the latter case the corrected correlation might be termed an index of *intrinsic validity*.

In order to obtain an experimentally unbiased estimate of a correlation corrected for attenuation, the instability errors must be the same in the intercorrelation(s) and the reliability coefficients. There appear to be only two experimental designs which will meet this requirement.

1. Administer both forms of both tests simultaneously. This requires the preparation of a special test booklet, with the items of the two tests (or of the two forms of the two tests) set up in a spiral-omnibus or cycle-omnibus arrangement. If either test consists of super items, the number of items in each super item will determine the length of the cycle. With this test format, the correlation corrected for attenuation is

$$r_{xy} = r_{xy} / \sqrt{C_x C_y},$$

with C_x and C_y the KR-20 consistency coefficients or the split-half coefficients raised by the Spearman-Brown formula.

2. Administer one form of each test at one sitting, and a parallel form of each test at another sitting at least a week later. Then

$$r_{xy} = \sqrt{\frac{r_{11I} r_{12}}{r_{11} r_{2II}}},$$

where 1 and I designate the two forms of one test, 2 and II the two forms of the other test, Arabic numerals the forms administered at one testing session, and Roman numerals the forms administered at the other session. Note that r_{12} and r_{1II} do not appear in this formula: every correlation which does appear is a correlation between a test administered at the first session and a test administered at the second session. It seems not unreasonable to assume that the instability errors in all pairs of tests administered a week or more

apart are essentially equal, even though at each session the forms are administered serially rather than simultaneously. Note also that the two forms of each test need not be equally reliable, equally variable, nor equally difficult. This formula, and the experimental design outlined above, are due to Yule, and appeared in an early edition of his *Introduction to the Theory of Statistics* almost 50 years ago.

To obtain an experimentally unbiased estimate of intrinsic validity, the administration of the two forms of the test or test battery should be spread over the whole interval during which the criterion behavior occurs. To do this we would require a large number of very short forms of the test, administered at random sub-intervals of the criterion-behavior period, and in most cases this would be administratively impractical. A second-best procedure would be to administer one form of the test or battery when the criterion-behavior period is one-fourth over and the other when it is three-fourths over. Each half-test would then consist of half the items of one form plus half the items of the other form, and the whole test would consist of the two forms combined. In this case,

$$r_{\infty} = r_{xy} / \sqrt{R_x R_y},$$

with r_{xy} the correlation between the whole test or battery and the whole criterion measure, R_x the correlation between the two half-tests (defined as above) raised by the Spearman-Brown formula, and R_y the correlation between the two half-criterion measures, also raised by the Spearman-Brown formula. This design requires that the two forms of the predictor test or battery and the two forms of the criterion measure be near enough equal in consistency and variability to meet the requirements of the Spearman-Brown formula. If they are not, let the half-tests of x be x_1 and x_2 , and let the half-tests of y be y_1 and y_2 . Then

$$r_{\infty} = \frac{\sqrt{r_{11} r_{22} r_{12} r_{21}}}{\sqrt{r_{11} r_{22}}}.$$

It could be argued that in the case of concurrent true validity, the test or battery should also be administered over the whole criterion-behavior period, or as a second-best in two parts at the end of one-fourth and three-fourths of the criterion-behavior period. The objective of concurrent true validity, however, is inconsistent with

such a procedure, since in practice the test or battery, as a substitute for the criterion-measurement procedure, would be administered all at one time.

Factor Analysis

When a number of tests which are to be factor-analyzed are administered serially, instability errors will attenuate the correlations between tests far apart in the series as compared with those administered consecutively or almost consecutively. There will also be a complex system of practice effects depending upon the particular order in which the tests are administered. There will be two main consequences: 1) some spurious augmentation of the general factor, due to the systematic lowering of correlations with increasing separation interval; 2) generation of a large number of small serial-effect error factors. With the usual sample sizes, these factors will not in general be larger than those generated by the sampling errors, but the two sets taken together may be sufficient to blur the effects of one or two of the smallest substantive factors, thus aggravating the problem of "when to stop factoring."

The best method of controlling these effects is again the method of simultaneous administration of all tests, using a spiral-omnibus or cycle-omnibus arrangement of the items in one or more special test booklets. If the battery is large, each booklet might contain only four or five items (or one superitem) from each test. A further refinement would consist in arranging the item sets in a different order in each booklet.

Five 40-item verbal tests were arranged in cycle-omnibus form in a single booklet (Cureton et al., 1944). Each cycle consisted of five items of each type (because the paragraph-reading test had five items on each paragraph), and there were eight cycles. The orders of the item types were not varied from cycle to cycle, though in theory this would have been advantageous. The tests were:

1. paragraph reading: paragraph followed by five four-choice questions;
2. same-opposite: two words to be identified as synonyms, opposites, class and member of the class, or none of these;
3. proverb matching: stem proverb and four answer proverbs; two to be identified as teaching the same lesson as the stem proverb (hence 6-choice items);

4. verbal analogies: two stem words followed by six alternative words; two to be *selected* and *arranged* to complete the analogy (hence 30-choice items);

5. vocabulary: stem word followed by five alternatives; one to be selected as the synonym of the stem word.

The test was administered without time limit to 841 high school students. Intercorrelations and split-half, Spearman-Brown reliabilities were computed, and the intercorrelations were factor-analyzed repeatedly by the centroid method until the communalities were stable $\pm .001$ from the beginning to the end of the last re-factoring. At the last re-factoring the numerically largest entry in the second residual matrix was .003. So far as the writer can recollect, he has not seen another final residual matrix in the literature whose numerically largest entry was as small as .03. This study therefore provides empirical verification of the expectation that simultaneous administration of the tests should improve the clarity of a factor analysis by removing small serial-effect factors.

The two factors were easily identified as a word-meaning factor (tests 2 and 5) and a verbal-reasoning factor (tests 1, 3, and 4). A further substantive finding of some interest was that the correlation corrected for attenuation between tests 2 and 5 was .981 (the next largest being .905, between tests 1 and 3). With this design, and using the correlations corrected for attenuation to supplement the results of the factor analysis, we can identify the 2:5 factor as an overlapping specific factor in contradistinction to an ordinary group factor. Tests 2 and 5, despite considerable differences in format, are in fact almost parallel forms.

If the spiral-omnibus or cycle-omnibus designs are impractical, and if there are two forms of each test, we could administer all the A forms in one session or series of sessions, and then some days later administer all the B forms. The rank correlation between the orders of administration of the A forms and the B forms should be zero or near-zero, subject to the further condition that no two tests administered consecutively at the first session should be administered consecutively (in either the same order or the reverse order) at the second session. This procedure would partially control the formation of practice-effect error factors. With this design there would be two usable correlations for each pair of tests: say r_{a1-b2} and r_{a2-b1} , where 1 and 2 refer to tests and a and b to forms. We would not use cor-

relations such as r_{a1-a2} or r_{b1-b2} , since these would be correlations between tests administered at the same session. Correlations such as r_{a1-b1} and r_{a2-b2} would be inter-form reliability coefficients experimentally fairly comparable to the intercorrelations if the interval between sessions were fairly large in comparison with the session lengths.

With data from such a design, we could assign each of the two correlations between any pair of tests randomly to one of two matrices, factor the two matrices separately, and obtain some empirical evidence concerning the inter-form reliabilities of the factors. The theory of factorial reliability from this design remains to be worked out, and so far as the writer is aware no such factor analysis has ever been performed.

"Coefficient" and "Index"

The literature of psychological and educational measurement is in many instances confused and incorrect because of failure to maintain consistently two distinctions: 1) the distinction between correlation coefficients (or standard deviation ratios) and coefficients of determination (or variance ratios); and 2) the distinction between defining formulas and computing or estimating formulas. The hoariest error of all is perhaps the dictum, current for 50-odd years, that "the upper limit of the validity of a test is the square root of its reliability."

It is well known that the square of any correlation coefficient, say r_{yx} , can be expressed as a variance ratio. If y is the dependent variable and x the independent variable, $r_{yx}^2 = \sigma_{y(x)}^2 / \sigma_y^2$, where for each observation $y(x) = b_{yx}x$, the distance from the y -mean to the corresponding point on the regression line for the given value of x ($= X - \bar{X}$). In like manner $r_{yx} = \sigma_{y(x)} / \sigma_y$, a standard deviation ratio. Also r_{yx}^2 is termed the coefficient of determination, because it is the proportion of the y -variance "determined by" (i.e., perfectly correlated with) x , while $k_{yx}^2 = 1 - r_{yx}^2 = \sigma_{y.e}^2 / \sigma_y^2$ is the coefficient of non-determination, the numerator of the last expression being the variance error of estimate.

In present standard statistical terminology we speak of the *coefficient* of determination and the *coefficient* of non-determination. This suggests that we should use the term "coefficient" to refer to a

variance ratio, leaving the term "index" to refer to a standard-deviation ratio.

Now in definitional language, the reliability coefficient (or the consistency coefficient) is a variance ratio: the variance ratio of true scores to raw scores. In reliability theory the dependent variable is the true score, and the independent variable is the raw score. The second (equally reliable) form is required only to *estimate* the value of this ratio, and in the case of consistency estimation no second form is required if the test is homogeneous in content and untimed in administration, and we use one of the Kuder-Richardson formulas. If x is a true score, x_1 and x_I the raw scores on two forms of the test, and R_1 and R_I the reliability (or consistency) coefficients of x_1 and x_I , then by definition,

$$R_1 = \sigma_x^2 / \sigma_1^2 = r_{x1}^2,$$

$$R_I = \sigma_x^2 / \sigma_I^2 = r_{xI}^2,$$

and if the two forms are equally reliable, we can *estimate* R_1 and R_I by

$$R_1 = R_I = r_{1I}.$$

Here r_{1I} is a computing or estimating formula, not a defining formula. We define the *index* of reliability (say r_1 or r_I) similarly:

$$r_1 = \sigma_x / \sigma_1 = r_{x1},$$

$$r_I = \sigma_x / \sigma_I = r_{xI},$$

and if the two forms are equally reliable, the *estimate* of r_1 and r_I is

$$r_1 = r_I = \sqrt{r_{1I}}.$$

The full argument, including in the first row the one given just above, is exhibited in Table 2.

From Table 2 it may be seen that if we are to maintain a consistent terminology we must re-name all of the "validity coefficients" in the literature as *indices of validity*. The "coefficient of alienation," moreover, $k_{yx} = \sqrt{1 - r_{yx}^2} = \sigma_{y \cdot x} / \sigma_y$, should be termed the *index of non-determination*, and to carry things to the extreme, the "correlation coefficient" itself would become the *index of determination*!

If we compare coefficients with coefficients or indices with indices, then as the reliability of a predictor approaches unity: 1) the upper

TABLE 2

*Defining and Computing Formulas for Coefficients
and Indices of Reliability and Validity*

General term	Defining formula		Computing formula	
	Coefficient	Index	Coefficient	Index
Reliability (or consistency)	σ_x^2/σ_1^2 or r_{x1}^2	σ_x/σ_1 or r_{x1}	r_{11}	$\sqrt{r_{11}}$
Raw validity	$\sigma_{o(1)}^2/\sigma_o^2$ or r_{o1}^2	$\sigma_{o(1)}/\sigma_o$ or r_{o1}	r_{o1}^2	r_{o1}
True validity	$\sigma_{v(1)}^2/\sigma_v^2$ or r_{v1}^2	$\sigma_{v(1)}/\sigma_v$ or r_{v1}	$\frac{r_{o1}r_{01}}{r_{o0}}$	$\sqrt{\frac{r_{o1}r_{01}}{r_{o0}}}$
Intrinsic validity	$\sigma_{v(x)}^2/\sigma_v^2$ or r_{vx}^2	$\sigma_{v(x)}/\sigma_v$ or r_{vx}	$\frac{r_{(o+o)(1+I)}}{\left(\frac{2r_{o0}}{1+r_{o0}}\right)\left(\frac{2r_{11}}{1+r_{11}}\right)}$	$\sqrt{\frac{r_{(o+o)(1+I)}}{\left(\frac{2r_{o0}}{1+r_{o0}}\right)\left(\frac{2r_{11}}{1+r_{11}}\right)}}$

limit of the raw validity is the criterion reliability, and 2) the upper limit of the true validity is the intrinsic validity. Also the upper limit of an actual test's raw validity is its reliability (when both its intrinsic validity and the criterion reliability are unity), and this is also the upper limit of its true validity. The upper limit of the intrinsic validity (or in general of any correlation corrected for attenuation) is unity; and this will actually occur whenever the test and the criterion (or the two tests) are essentially parallel forms.

When (as happens) we see in the literature a correlation corrected for attenuation which is substantially greater than unity, we can usually assume that the formula used for its computation was inconsistent with the experimental design. If, for example, the formula,

$$r_{\infty o} = \frac{\sqrt{r_{12}r_{11}r_{12}r_{11}}}{\sqrt{r_{11}r_{211}}}$$

were used with the two-session design, which calls for Yule's formula, the value of $r_{\infty o}$ would be substantially overestimated, since r_{12} and r_{11} are same-session correlations while the denominator contains only intersession reliability coefficients. Essentially the same effect would be given by

$$r_{\infty} = \frac{r_{(1+I)(2+II)}}{\sqrt{\left(\frac{2r_{1I}}{1+r_{1I}}\right)\left(\frac{2r_{2II}}{1+r_{2II}}\right)}}$$

These formulas apply properly only when both forms of each test are administered simultaneously. The last was in fact the one used in the study of five verbal tests mentioned above.

Summary

1. The basic assumptions of reliability and validity theory (errors of measurement uncorrelated with one another and with true scores) imply definite restrictions on test-research experimental designs. In most cases the crucial point is the avoidance of correlated time-associated errors.

2. *Inconsistency* errors are associated with forms; *instability* errors with times of administration. The *stability coefficient* is the inter-form correlation corrected for attenuation due to the inconsistencies of the two forms.

3. Test *consistency* is measured by α (KR-20) or by the split-half method and the Spearman-Brown formula. In the split-half method, the two half-tests must be administered simultaneously. All items relating to a single exhibit (e.g., paragraph or table) must be assigned to the same half-test, or treated as one superitem in computing α . The consistency of a speed test may be estimated by Feldstein's method if at least five equivalent forms have been administered at equal intervals.

4. Instability increases continuously with the time between the administration of two forms of a test. Stability coefficients and inter-form reliability coefficients depend explicitly on the time interval. For intervals on the order of a week or two, stability coefficients are numerically of the same order of magnitude as consistency coefficients based on 30- or 40-item forms, and inter-form reliability coefficients are appreciably lower.

5. The Spearman-Brown formula applies only to forms or half-tests administered simultaneously. Lengthening a test increases its consistency but has no effect upon its stability.

6. Raw validity is the correlation between a predictor and a *sui generis* criterion measure.

7. True validity is the correlation between a predictor and the

estimated true scores on a constructed criterion measure: a predictor-criterion correlation corrected for criterion attenuation.

8. Intrinsic validity is the correlation between estimated true scores on a predictor and estimated true scores on a criterion measure: a predictor-criterion correlation corrected for both predictor and criterion attenuation.

9. In every formula for correction for attenuation, the time interval between the administration of the two tests or forms must be the *same* interval for every correlation coefficient entering into the formula. Two experimental designs are suggested to accomplish this.

10. In factor analysis serial (as against simultaneous) administration of the tests increases the size of the general factor and introduces small serial practice effect factors.

11. To resolve long-standing confusions, we should distinguish sharply between defining formulas and computing formulas, and between indices (correlations or standard deviation ratios) and coefficients (squared correlations or variance ratios). The main confusion comes from the fact that the reliability coefficient is by definition the *square* of the correlation between the raw score on one form and the true score on that form, and is computed as the *un-squared* correlation between the given form and an equivalent form.

REFERENCES

- Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, XVI (1951), 297-334.
- Cureton, E. E., et al. "Verbal Abilities Experiment: Analysis of New Word Meaning and Verbal Analogies Tests." PRS Report No. 548, Personnel Research Section, The Adjutant General's Office, War Department, 1944 (mimeographed).
- Cureton, E. E. "The Definition and Estimation of Test Reliability." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVIII (1958), 715-738.
- Loveland, E. H. "Measurement of Factors Affecting Test-Retest Reliability." Unpublished Ph.D. Thesis, University of Tennessee, 1952.
- Newstetter, W. I., Feldstein, M. J., and Newcomb, T. M. *Group Adjustment: A Study in Experimental Sociology*. Cleveland, School of Applied Social Sciences, Western Reserve University, 1938.
- Taylor, E. K., Manson, Grace E., and Stone, P. M. "Validation of Tests for Routine IBM Jobs at Office of Dependency Benefits, Newark, New Jersey." PRS Report No. 698, Personnel Research Section, The Adjutant General's Office, War Department, 1945 (mimeographed).

THE EFFECTS OF PARTIAL-PACING ON TEST PARAMETERS¹

JAMES M. ELLIOTT AND H. G. OSBURN
University of Houston

It is generally recognized that individuals have different rates of work, and there are intra-individual variations of work rates across time. In the area of psychological testing, these variations may contribute to the error variance component of the measure under investigation, and consequently reduce the reliability of the test score (Walker and Lev, 1953; Cronbach, 1960). This study is concerned with those individual differences in work rate which may be experimentally controlled by the examiner. One form of such control is pacing. Pacing is controlling the time allowed for a response to the stimulus, i.e., test item. A completely paced test is one in which time is controlled for each item; a partially-paced test is one in which several short series of items are presented with the time for each series of items being proportional to the total working time of the original unpaced test. An unpaced test is one in which examinees are permitted to attempt as many items as possible in the time allotted for the entire test.

Sax and Carr (1962) rearranged items on a well-known spiral omnibus intelligence test to form two homogeneous subtests. Subtests were administered as separately timed tests and compared with scores obtained on an unaltered form of the same test. They found that people attempted significantly more items on unaltered forms than on altered ones, and reliability of the altered form was signifi-

¹ This paper has been modified from a thesis submitted by the senior author to the Department of Psychology, University of Houston in partial fulfillment for the degree Master of Arts (1963).

cantly lower (p. 372). However, upon correlating scores from both unaltered and altered tests with grade-point average, they found no difference in validity (p. 375). Thus Sax and Carr found a lower work rate under partial pacing but the results were obscured by differences in difficulty level of the two altered forms.

Myers (1952) presented some data which suggests that variable work rates are a function of the number of items and the time allowed to work on the test. A figure classification test was administered in five 12-minute parts, with 10, 20, or 30 items per part. Data showing percentage of people completing each part tend to lend supportive evidence for variable work rates within a test. In general, he found that keeping testing time constant and increasing the number of items in the test resulted in an increase in the number of items attempted per unit time. He concluded, however, that moderately speeded tests were more valid in predicting the criteria, i.e., grades and grade averages.

Agnoff (1953), in a methodological study concerning effective length of a test, compared a short test with two longer ones. He concluded the shorter test had greater "effective length," than the two longer ones, greater than would be predicted from the relative number of items in the tests (Agnoff, 1953, p. 13).

It was predicted that partial pacing would significantly increase the number of items attempted on a relatively pure speed test; and there would be significantly more items wrong on the partial paced tests (due to attempting more items in the time allotted). Higher reliability estimates were predicted for the partial paced test due to additional experimental control over test conditions, and as a corollary, correlations between the partial paced tests should be higher due to reduction of error variance.

Method

Subjects

Subjects were University of Houston undergraduate students enrolled in introductory psychology courses. Complete test data were obtained on 47 students tested under normal testing procedures (not paced), and 48 students tested under partial paced conditions. Six different sections were tested. Three sections were randomly assigned to paced conditions, and three to unpaced conditions. No sig-

nificant differences were found between mean test scores of the three sections within the unpaced conditions or within paced conditions.

Description of the Tests

Three homogeneous tests were used in this study. In each test the task to be performed was intellectually rather simple and highly repetitive. Two forms of each test were constructed. Each form was further divided into an unpaced and partial paced version. Thus there were four versions of each test: Form A-unpaced, Form A-paced, Form B-unpaced, Form B-paced; each version containing 50 items. Partial paced versions were constructed so that only 10 items appeared on a page.

Tests were as follows.

(1) *Clerical Carefulness Test BE-464AX*. This test was developed by the Human Resources Research Center, Lackland Air Force Base, Texas. Each item consists of a row of 15 3-digit numbers; the examinee's task is to pick out the highest number in the row. The reliability (odd-even) of the original test is estimated to be .88 (Osburn, Sheer, Elliott, and Mullins, 1963).

(2) *Letter Counting Test PL 0092*. This test was developed by the Personnel Laboratory, Lackland Air Force Base, Texas. Each item consists of a lengthy series of randomly arranged letters. The examinee is to count the number of "e's" appearing in each series of letters. The estimated reliability (odd-even) of this test is .88 (Osburn, *et al.*, 1963).

(3) *Letter Matrix Test*. This test was developed by Elliott (Osburn, *et al.*, 1963). The examinee is required to perform a sequential operation (analogous to multiplication) on letters. The estimated reliability (odd-even) is .88 (Osburn, *et al.*, 1963).

Test Administration

Tests were administered in two successive class periods. During the first class period, Form A of each test was administered, and the following class period Form B was given. Tests were given in the same order each time for both unpaced and partial paced versions. The Letter Matrix Test was administered first, Letter Counting Test second, and Clerical Carefulness Test third.

The time limit for Letter Counting and Clerical Carefulness tests

was 10 minutes total "working" time; time limit for the Letter Matrix test was 15 minutes. Thus, for the partial paced versions of the Letter Counting and Clerical Carefulness tests, the time limit for each page (10 items) was two minutes, and for the Letter Matrix test, the time limit was two and one-half minutes per page (10 items).

After test directions were read and before work was begun on the tests, subjects were informed of the time they would be allowed to work on the tests; or in the case of the partial paced versions, how long they would be allowed for each page.

Results

Effect of Partial Pacing on Number of Items Attempted, Number Right, and Number Wrong

The two forms of each test were combined for these analyses, and results are shown in Table 1. Mean number of items attempted was significantly different between the partial paced and unpaced conditions for all three tests and in each instance more items were attempted by the partial paced group. These data substantiate the

TABLE 1

Means and Standard Deviations of the Number of Items Attempted, Number of Items Wrong, and Number of Items Right on Both Versions of the Three Tests

	Paced Mean	Unpaced Mean	<i>t</i>	Paced S.D.	Unpaced S.D.	<i>F</i>
Letter Matrix						
Attempted	70.98	58.36	4.04**	14.20	15.87	1.25
Wrong	13.06	9.34	2.28*	9.33	6.15	2.30**
Right	57.91	49.02	2.54**	17.40	16.33	1.14
Letter Counting						
Attempted	95.12	82.45	7.04**	6.51	12.84	3.87**
Wrong	14.00	13.43	.35	8.29	8.55	1.06
Right	81.12	69.02	4.64**	11.01	14.01	1.62
Clerical Carefulness						
Attempted	97.96	88.76	5.21**	3.31	11.69	12.46**
Wrong	12.85	11.62	.88	6.54	6.93	1.12
Right	85.10	77.13	3.83**	7.69	12.06	2.46**

*Significant at the .05 level of confidence.

**Significant at the .01 level of confidence.

hypothesis that subjects will attempt more items under partial paced conditions than under unpaced conditions. In addition it was found that the standard deviation of the number of items attempted was lower under partial paced conditions, and the differences were sta-

tistically significant for the Letter Counting and Clerical Carefulness tests. This finding suggests that subjects taking the partial paced tests adopted a more uniform, as well as a higher, work rate as compared to subjects taking the unpaced tests.

Mean number of items wrong differed significantly only on the Letter Matrix test. Mean wrongs scores for the other two tests were essentially the same for both conditions. It was predicted that there would be more items wrong on the partial paced tests due to the increased work rate under the partial paced condition. However, this was generally not the case. It would appear that subjects attempted more items on the partial paced Letter Counting and Clerical Carefulness Tests without increasing their wrongs score. Apparently the reasons that more errors were made on the Letter Matrix test under the partially paced condition were due to the fact that this test is the most complex of the three, and the time limit imposed on each set prevented all but the very fastest subjects from checking their answers.

As for the number of items right, the means are significantly different, with the means being larger for the partial paced condition on each test. It appears that the consistent effect of increasing the work rate by partial pacing on the type of test utilized in this study was to increase the number of items attempted, and most of the additional items were answered correctly.

Effect of Partial Pacing on Reliability Estimates

Table 2 shows parallel forms reliability estimates and intercorrelations among the three tests. Product-moment correlation co-

TABLE 2

Reliability Estimates, Standard Errors of Measurement, and Intercorrelations for the Partial Paced and Unpaced Versions of the Three Tests

Test	Condition	Reliability	Standard Error of Measurement		Intercorrelations	
			Form A	Form B	Test 2	Test 3
Letter Matrix	Paced	.78	4.14	4.53	.44	.50
	Not Paced	.80	3.80	3.90	.69	.59
Letter Counting	Paced	.72	3.36	2.92	—	.57
	Not Paced	.76	3.85	3.44	—	.66
Clerical Carefulness	Paced	.60	2.66	2.77	—	—
	Not Paced	.73	3.22	3.44	—	—

efficients were computed using the rights score only. The reliability estimates are slightly lower for the partial paced version of all three tests, but the differences are not statistically significant. Pacing does not improve the reliability of these tests, and probably has the effect of slightly lowering reliabilities.

Effect of Partial Pacing on Interrelations among Tests

Intercorrelations between tests also tend to be somewhat lower for the partial paced tests. Again, however, differences between the pairs of entries in Table 2 are not statistically significant. As in the case of the reliability estimates, pacing reduces the correlations somewhat.

To further study effects of partial pacing on test reliability and error measurement, the standard error of measurement was computed for both forms of each test. These data are shown in Table 2. Except for the Letter Matrix test, standard errors of measurement were lower on partial paced tests even though parallel-forms reliability estimates were slightly lower under the same condition.

Discussion

The results suggest that partial pacing does have considerable effect on certain test statistics for the type of speed test used in this study. Specifically, subjects taking the partial paced test attempted significantly more items than subjects taking the same test under unpaced conditions. Moreover, as a result of having attempted more items, the partial paced group made significantly higher scores as compared to the unpaced group.

Except for the Letter Matrix test, wrongs were not increased appreciably by partial pacing. Since the probability of making an error on any given item in the Letter Counting and Clerical Carefulness tests is rather low, it would appear that encouraging subjects to respond at a faster rate did not materially increase the probability of making an error on a particular item. Thus, the subject's score depends on how far he was able to go in the time allowed. Under partial paced conditions, the subjects are encouraged "to go farther" in the test, and consequently, significantly increase their scores on the number right without a commensurate increase in errors.

It seems clear that partial pacing introduces a rather uniform response set for speed. Cronbach (1950) has suggested that response

sets tend to reduce the range of individual differences. Generally, our data lend supportive evidence for this hypothesis. The reduction in the range of individual differences due to partial pacing is particularly evident in number of items attempted, though not so clear in number right or number wrong.

Reliability estimates are not significantly different under the two conditions, although they are consistently higher on the unpaced tests. On the other hand, the standard error of measurement tends to be lower for the partial paced tests, although again, they are not significantly different. Thus, both error variance and systematic variance tends to be reduced by partial pacing. By having reduced the error variance, the tests were improved to a degree, and one would expect that reliability estimates would be higher. However, work-rate component of true variance was attenuated to such an extent that reliabilities were actually lower under partially paced conditions even though there was a slight reduction in error variance.

The introduction of a speed-set in the partially paced tests also probably accounts for the somewhat lower intercorrelations among these tests as compared to unpaced versions. Some of the common variance in the unpaced tests is due to differential work-rates, thus a person's true score on the unpaced versions is not only composed of ability, but also involves an individual work-rate component. Partial pacing tends to cancel out these differential work-rates by encouraging all subjects to work at relatively high rates.

REFERENCES

- Agnoff, W. H. "Test Reliability and Effective Test Length." *Psychometrika*, XVIII (1953), 1-14.
- Cronbach, L. J. "Further Evidence of Response-Set and Test Design." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, X (1950), 3-31.
- Cronbach, L. J. *Essentials of Psychological Testing*. (Second Edition). New York: Harper Brothers, 1960.
- Myers, C. T. "The Factorial Composition and Validity of Differentially Speeded Tests." *Psychometrika*, XVII (1952), 347-352.
- Osburn, H. G., Sheer, D. E., Elliott, J. M., and Mullins, C. J. "The Construction and Validation of a Battery of Carefulness Tests against School Criteria." Unpublished research report. University of Houston, 1963.
- Sax, G. and Carr, A. "An Investigation of Response Sets on Altered Parallel Forms." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 371-376.
- Walker, Helen M. and Lev, J. *Statistical Inference*. New York: Henry Holt & Co., 1953.



RISK TAKING AND ACADEMIC SUCCESS AND THEIR RELATION TO AN OBJECTIVE MEASURE OF ACHIEVEMENT MOTIVATION¹

ALBERT E. MYERS²

Yale University

If given a choice, most psychologists would probably prefer to use an objective test in preference to a projective one if it were guaranteed that the tests had identical reliabilities and validities. In research to date on *n* Achievement, the part played by objective tests has not been very impressive. To begin, the great bulk of the significant work done in this area has been done with varied objective procedures, e.g., Atkinson (1958a), McClelland (1955; 1961), McClelland et al. (1953). Further, studies which have used objective tests, e.g., Atkinson and Litwin (1960), Barnette (1961), Charms, Morrison, Reitman, and McClelland (1955), Izard (1962), have generally not been able to produce the same results. Thus far, it would seem that McClelland's (1958a) argument, that *n* Achievement can only be measured with projective techniques, is difficult to refute.

This paper, on the other hand, will report the highly satisfactory use of a short objective test. The test was used in two studies which used high school students. One was a correlational study concerned

¹ This study was supported in part by funds from the Office of Naval Research, Contract Nonr 2959(00). I would like to express my appreciation to Rodney W. Skager, Charles B. Schultz, and Anne M. Bussis for their generous consent to report original data that has not been previously published and for their comments on the manuscript. My appreciation also goes to the readers of earlier drafts, Thomas L. Hilton and Nathan Kogan.

² Formerly at the Educational Testing Service.

with academic achievement and the other was a laboratory experiment concerned with competition and risk taking.

Method

The items for the Achievement Motivation scale were included as part of a battery of tests used in a study of academic achievement and guidance. Partial results from this project have been reported by Skager, Bussis, and Schultz (1963) and Schultz and Skager (1963). Most of the items were selected from the Personal Values Inventory by Schlessler. Without resorting to any particular theoretical framework, 10 items were included which were aimed directly at academic achievement. No attempt was made to form an *n* Achievement scale, i.e., a measure of the individual's generalized need for achievement. The intent was to determine if the students were trying to achieve in school. The items were the following.

1. When you know there are going to be one or two questions on an exam from outside reading assignments, do you always read *all* the material?
2. Do you regard yourself as a more consistent and harder worker in your classroom assignments than the typical high school student in your classes?
3. Have others (not your good friends) thought of you as one who "missed some of the fun" because you were so serious?
4. Do you think your fellow students in high school think of you as a hard worker?
5. Do most of your teachers probably think of you as one of their hardest workers even though not necessarily one of the brightest?
6. Do other interests (sports, extra-curricular activities, or hobbies) prevent you from obtaining an excellent rating or mark for *effort* in school work?
7. Do you have a very strong desire to excel academically?
8. Do you try harder to get on the school honor roll or merit list than the average student in your class?
9. Do you try to do most jobs at least a little better than what you think is expected?
10. Do you tend to give up or delay on uninteresting assignments?

11. Which do failures most often tend to do to you?

(Y) Start you off on some new interest.

(N) Spur you to new efforts in the thing at which you failed.

12. Are your friends more likely to consider you as

(Y) casual and carefree?

(N) responsible?

The Ss were required to answer "yes," "no," or "?" ("can't say," "doesn't apply"). The battery was administered to 261 male and 263 female high school juniors. The students were members of "college bound" groups from seven eastern high schools. The mean PSAT-Verbal and PSAT-Math scores for males were 48.5 and 55.5 while the mean PSAT-V and PSAT-M were 47.3 and 47.6 for females.

In a subsequent laboratory study (Myers, 1964) the same scale, with some alteration, was used again. The last three items were dropped and the Ss were required to indicate the degree to which the statement was true on a six-point scale.

The Myers study was concerned with the acquisition of psychological advantages in competitive situations. Twenty high school students who were paid for their time served as subjects. The Ss played in a shuffleboard bowling tournament. Unknown to the Ss the apparatus (a commercial "toy" frequently found in amusement centers) had been wired so that the performance of the players was controlled by the experimenter. The Ss had virtually no influence on the determination of their performance scores.

The tournament was designed in a way which permitted measurements of risk taking and optimism. Risk was determined by the number of points an S tried to win in a game. If an S gave his opponent a large handicap and thereby lowered the probability of winning, he was rewarded with more points (if he won) than if he gave a small handicap. In short, risk varied directly with payoff.

The optimism score was based upon predictions made by the S of the scores that he and his opponents would get in the next game. If the predictions indicated that the S thought he was going to do well in relation to his opponent, a positive (or optimistic) score was given. If, on the other hand, the S predicted he would do poorly in relation to his opponent, a negative (or pessimistic) score was given.

Results

Motivation, Intellectual Ability, and Academic Achievement.

Table 1 shows the correlations between the motivation scale, grade-

TABLE 1
Intercorrelations of Achievement Motivation, PSAT
and Grade-Point Average for Males and Females^a

	PSAT-V	PSAT-M	GPA
Achievement Motivation	.16* (.13)	.21 (.19)	.50 (.48)
PSAT-V		.71 (.60)	.61 (.55)
PSAT-M			.59 (.62)

^aValues for females are given in parentheses.

* $r_{.05} = .13$; $r_{.01} = .16$; $r_{.001} = .21$.

point average, the Preliminary Scholastic Aptitude Tests for verbal (PSAT-V) and math (PSAT-M) that were found in the project by Skager, Schultz, and Bussis. The correlation of the Achievement Motivation score with grade-point average was .50 and .48 for males and females respectively. This compares quite favorably with the .51 found by McClelland et al. (1953, p. 237) using TAT pictures, the .19 to .36 found by Ricciuti and Sadacca (1955) using the TAT, the .32 to .38 found by Barnette (1961), Bendig and Klugh (1956), and Gough (1953) using objective devices. Every bit as interesting is the fact that the correlations of achievement motivation and the measures of intellectual ability range between .13 and .20. This is lower than any similar correlation found in the above studies. It appears, therefore, that the present scale had a greater incremental validity due to motivation when combined with intelligence than any of the previously reported scales. It seems safe to conclude that the present scale adequately reproduced the findings, with respect to academic achievement, that had previously been found with projective devices.

Motivation and Risk Taking. It has frequently been found with projective devices that a curvilinear relationship exists between achievement motivation and risk taking, with high achievement motivation being associated with intermediate risk (Atkinson, 1958b; Atkinson, Bastian, Earl, and Litwin, 1962; Atkinson and Litwin, 1960; Litwin, 1959; McClelland, 1958b). The *n* Achieve-

ment scale designed by Edwards for the Personal Preference Schedule was not able to reproduce this result.

Figure 1 shows a plot of points which describes the relationship between motivation and risk taking as found in the shuffleboard bowling experiment. The risk taking scores represent the number of points tried for over a period of games corrected for differences in experimental conditions. Constants have been added to both the risk taking and motivation scores to avoid negatives. The relationship was curvilinear. Using four categories on both the independent and dependent variables, a correlation ratio of .75, which was significant at the .01 level, was obtained. Since the absolute magnitude of the correlation ratio varies with the number of categories used in determining it (as does χ^2), it is difficult to say what the correlation ratio was. But no matter how many categories greater than three were used, the correlation ratio was highly significant. The traditional finding between achievement motivation and risk taking was, therefore, adequately reproduced.

Motivation and Optimism. Feather (1963a, 1963b) hypothesized that there should be a positive relationship between *n* Achievement and "expectation of success." He was not able to support that hypothesis with his data. Similarly, that hypothesis was not supported with the present data. The correlation between motivation and optimism was not even significant at the .10 level.

Discussion

It does not seem necessary to argue that the scale used in these studies was a satisfactory substitute for the projective measures of *n* Achievement. The data speak for themselves. But, unfortunately, they do not tell us why this scale performed so well.

Proponents of the projective approach to *n* Achievement will undoubtedly regard these results with considerable skepticism—and with good reason. Numerous attempts have been made since 1953 to make a satisfactory objective scale and all have met with failure. Why should this alarmingly short scale be so effective? The answer is not at all clear, but certain things may be taken into account.

It is always possible that any set of findings may be the result of chance. That could hardly be true in the present case. The probability of reproducing all these results at the observed levels of significance is truly remote.

ACHIEVEMENT MOTIVATION

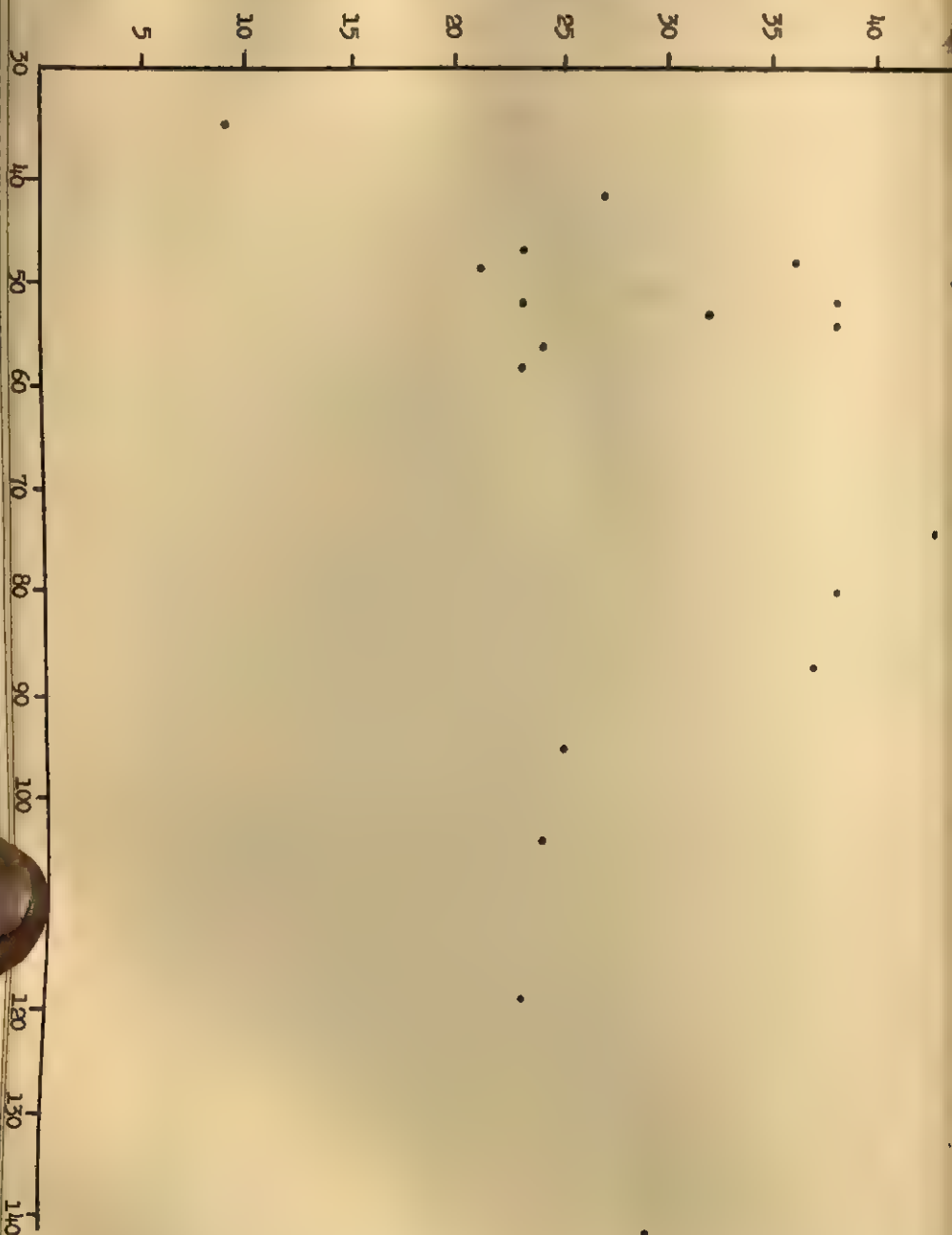


Fig. 1. Relationship between achievement motivation and risk taking.

Perhaps the value of the scale lies in its utter simplicity. Instead of asking such profound and soul-searching questions as "I feel that my future peace and self respect depend upon my accomplishing some notable piece of work" (deCharms et al., 1955), or "Only a fool would try to change our American way of life" (Gough, 1953), the present scale essentially asks "Do you try to do well in school?" Not so oddly, students who seem to have a higher need to achieve answer "yes."

The test, of course, has some obvious drawbacks. Since it could easily be faked it would probably be quite unacceptable as a selection device. It is also possible that the scale would be less effective on a non-school population, thereby reducing its value as a research instrument. The scale in its present form is certainly susceptible to the influences of acquiescence and social desirability. There is no guarantee, in fact, that the scale is even measuring motivation, in any form. All that can be asserted at present is that the scale was able to reproduce the findings typically found with projective devices with college-oriented high school students. Despite these drawbacks, it does seem possible that we may hold out hope for the development of objective measures of *n* Achievement. There are, at last, some data in which objective scores do just as well as projective ones.

Summary

Two studies using an objective test of Achievement Motivation with high school Ss indicated that the test was able to reproduce the results typically found with projective measures. One was a correlational study dealing with problems of academic achievement while the other was a laboratory study that focused on competitive risk taking. These data seemed to support the proposition that objective tests of achievement motivation may be possible.

REFERENCES

- Atkinson, J. W. (Ed.) *Motives in Fantasy, Action and Society*. Princeton, N. J.: Van Nostrand, 1958. (a)
Atkinson, J. W. "Toward Experimental Analysis of Human Motivation in Terms of Motives, Expectancies and Incentives." In J. W. Atkinson (Ed.), *Motives in Fantasy, Action and Society*. Princeton, N. J.: Van Nostrand, 1958. (b)
Atkinson, J. W., Bastian, J. R., Earl, R. W., and Litwin, G. H. "The Achievement Motive, Goal Setting and Probability Prefer-

- ences." *Journal of Abnormal and Social Psychology*, LX (1962), 27-36.
- Atkinson, J. W. and Litwin, G. W. "Achievement Motive and Test Anxiety Conceived as Motive to Approach Success and Avoid Failure. *Journal of Abnormal and Social Psychology*, LX (1960), 52-63.
- Barnette, W. L., Jr. "A Structured and Semi-structured Achievement Measure Applied to a College Sample." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 647-656.
- Bendig, A. W. and Klugh, H. E. "A Validation of Gough's Scale in Predicting Academic Achievement." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVI (1956), 516-523.
- deCharms, R., Morrison, H. W., Reitman, W., and McClelland, D. C. "Behavior Correlates of Directly and Indirectly Measured Achievement Motivation." In D. McClelland (Ed.), *Studies in Motivation*. New York: Appleton-Century-Crofts, 1955.
- Feather, N. T. "The Relationship of Expectation of Success to Reported Probability, Task Structure, and Achievement Related Motivation." *Journal of Abnormal and Social Psychology*, LXVI (1963), 231-238. (a)
- Feather, N. T. "The Effect of Differential Failure on Expectation of Success, Reported Anxiety and Response Uncertainty." *Journal of Personality*, XXXI (1963), 289-312. (b)
- Gough, H. G. "The Construction of a Personality Scale to Predict Scholastic Achievement." *Journal of Applied Psychology*, XXXVII (1953), 361-366.
- Izard, C. E. "Personality Characteristics (EPPS), Level of Expectation, and Performance." *Journal of Consulting Psychology*, XXVI (1962), 394.
- Litwin, G. H. "Achievement Motivation, Social Class, and the Slope of Occupational Preferences in the United States and Japan." Dept. of Social Relations, Harvard University, 1959. (Dittoed paper.)
- McClelland, D. C. *Studies in Motivation*. New York: Appleton-Century-Crofts, 1955.
- McClelland, D. C. "Methods of Measuring Human Motivation." In J. W. Atkinson (Ed.), *Motives in Fantasy, Action and Society*. Princeton, N. J.: Van Nostrand, 1958. (a)
- McClelland, D. C. "Risk Taking in Children with High and Low Need for Achievement." In J. W. Atkinson (Ed.), *Motives in Fantasy, Action and Society*. Princeton, N. J.: Van Nostrand, 1958. (b)
- McClelland, D. C. *The Achieving Society*. Princeton, N. J.: Van Nostrand, 1961.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., and Lowell, E. L. *The Achievement Motive*. New York: Appleton-Century-Crofts, 1953.
- Myers, A. E. "Performance Factors Contributing to the Acquisition of a Psychological Advantage in Competition." *Human Relations*, in press.

- Ricciuti, H. N. and Sadacca, R. "The Prediction of Academic Grades with a Projective Test of Achievement Motivation: II. Cross-validation at the High School Level." Research Bulletin 55-16. Princeton, N. J.: Educational Testing Service, 1955. (Multilithed Report.)
- Schultz, C. B. and Skager, R. W. "Relationship of an Independent Activities Questionnaire to Performance during High School." Research Bulletin 63-16. Princeton, N. J.: Educational Testing Service, 1963. (Multilithed Report.)
- Skager, R. W., Bussis, Anne M., and Schultz, C. B. "Comparison of Information Scales and Like-Indifferent-Dislike Scales as Measures of Interest." Research Bulletin 63-10. Princeton, N. J.: Educational Testing Service, 1963. (Multilithed Report.)



16 PF ITEM RESPONSE PATTERNS AS A FUNCTION OF REPEATED TESTING^{1, 2}

KENNETH I. HOWARD

AND

HERMAN DIESENHAUS

Institute for Juvenile Research

In a previous paper (Howard, 1964) analyses were reported of the effects of repeated testing on the differentiation of individuals. Using profile correlations and item agreement scores for a variety of tests and samples, it was found that first test scores were less reliable and less differentiating than were later test scores. Increasing consistency of *individual* test performance was demonstrated when the agreement between later pairs of occasions was greater than agreement between an individual's first and second test performances. Increasing stability of test results was also indicated by higher test-retest correlations for *scale* scores on later pairs of occasions than on the first pair. Increasing differentiation of individuals was demonstrated by a decrease in the correlations between subjects on later occasions. These findings led to the conclusion that first test scale scores are psychometrically inferior to scores obtained on later administrations of the same test, and that repeated testing would serve to provide more stable and unique data on which to base descriptions of an individual personality.

This paper reports analyses of personality inventory item-

¹ This study was supported in part by a grant from the Psychiatric Training and Research Fund of the Illinois Department of Public Welfare. The authors also wish to express their appreciation for assistance of Mr. Wil Blair, Miss Hannah Frisch, and Miss Shirley Breslow.

² Extended version of paper presented at the Fourth Annual Meeting of the Society of Multivariate Experimental Psychology, Boulder, 1963.

response characteristics in order to further document these repeated testing effects, extending the generality of the phenomena from the level of test scale scores to that of items. This involves a shift in focus from emphasis on agreement between persons on the same and subsequent test occasions to an emphasis on the consistency of responses to single items or classes of items.

Previous analyses of word association test data (a free response situation) had shown that with repeated administrations of test stimuli, there was a tendency toward decrease in the frequency of "common" or popular responses (as defined by the proportion of subjects giving a response on the first trial) while there was an increase in the number of stable unique responses (Howard, 1964). Such a decrease in popular responses and increase in stable unique responses to single stimuli represents better differentiation of individuals on later occasions. Analysis of repeated administration of the MMPI (a limited response situation) showed that there was a tendency toward a 50-50 split over repeated trials for the endorsement frequency of items with two response alternatives. Of course, a 50-50 response split would give maximum differentiation of individuals. This increase of item response variance also represented a decrease in endorsement of popular responses and an increase in differentiation of individuals in the repeated two-choice situation.

From a somewhat different point of view, there has been increasing interest in the role of structural characteristics of individual items, other than the specific content purporting to measure some need or trait, as a determinant of response tendencies (Buss, 1959; Cronbach, 1946; Edwards and Walsh, 1963; Fiske, 1957a, 1957b, 1961; Goldberg, 1963; Hanley, 1959; Wiener, 1948; Wiggins, 1962). Various structural characteristics or "item parameters" (Edwards and Walsh, 1963) have been identified as affecting the probability of initial endorsement and the stability of personality inventory responses. Previous analyses have utilized, among other characteristics, number of response alternatives, item ambiguity, item variance, social desirability scale value, and item wording. Another obvious item characteristic is the difference in wording of response alternatives. The relationships between endorsement frequency, response stability, and response format have not previously been studied. By classifying items according to the type of response format and the wording of response alternatives, and analyzing the

retest effects for each format, it will be possible to obtain a clear analysis of trends (appearing when there is no relevant experience intervening between consecutive administrations of the same test) and to extend findings about item structure.

Methods and Procedures

Subjects

The 80 subjects were students in an out-service training program sponsored by the Illinois State Employment Service and offered by the University of Chicago. All of the subjects were professional employment counselors. There were 43 women and 37 men ranging in age from 23 to 65 years.

Procedure

The subjects were administered Form A of the 16PF (Cattell *et al.*, 1957) once a week, during class sessions, for three weeks. There were four groups of approximately 20 students each. In all cases the testing interval was exactly one week, and all the subjects were tested at approximately 9:30 in the morning. (Although trait variability would be expected, the constant testing interval provided an adequate control of this source of variance for the purposes of analysis of repeated testing effects.) The standard instructions were read each time, but on the second and third testing occasions the subjects were cautioned:

You should all recognize this test. It is the same one you took last week. As you go through the test you will probably remember some of your previous answers. Try not to let that influence the answer you give today. I do not want you to try to give the same answers as you did last week, or to try to change your answers. Just give your first, natural, honest answer to each question on the basis of what is true today.

In most cases the subjects did not know how many times they were to take the test. The purpose of the testing was not divulged until after the last session.

Instrument

The 16PF is a personality questionnaire consisting of 187 questions, each with three response alternatives. This test was well suited

for the purposes of the study since a variety of formats are employed to measure the different traits. Thirteen questions are intelligence items with "right" answers, and three questions are essentially instructional. With the exception of three additional items, the remaining items could be grouped into the following five major response formats.

- I. "Yes - In Between - No"—84 items were in this category. Included were all Yes-No, True-False, and Agree-Disagree items with the third response labelled "In Between."
- II. "A - In Between - C"—24 items were choice items with an "In Between" response. An example of a choice item is "I prefer to marry someone who: (a) commands a general admiration, (b) in between, (c) has artistic and literary gifts."
- III. "Yes - Uncertain - No"—11 items of the same general type as Format I contained a third response labelled "Uncertain."
- IV. "A - Uncertain - C"—23 items of the same general type as Format II contained a third response labelled "Uncertain."
- V. "Continuum"—26 items contained three responses defining a continuum. An example of a Format V item is, "I can find enough energy to face my difficulties. (a) always, (b) generally, (c) seldom."

Results

All analyses were carried out for each format separately. In general, statistical tests of the results were not undertaken, nor were confidence statements regarding estimates of parameters. Rather the interpretations rest on the pattern of results and consistency over formats.

Since there were three possible responses to each item, and three testing occasions, there were 27 possible item response patterns. Frequency counts of these 27 patterns were accomplished for all items, and for each format. Rather than using the original response labels ("a," "b," "c"), the responses were coded X, B, Y. All "b" responses (e.g., "In Between" or "Uncertain") were coded B. Of the remaining responses ("a" and "c") the more popular response on the first testing occasions was coded X, and the least popular response was coded Y. Table 1 reports the basic data of this study.

For ease of presentation, frequency counts have been converted to proportions. Thus, P(XXX) represents the proportion of responses for which the popular response to an item was endorsed on all three occasions. P(BYY) represents the proportion for which the B response was endorsed on the first occasion, and the minority response was endorsed on the next two occasions.

TABLE 1

Proportion of Responses in Each Possible Response Pattern for Each of the Five Basic Response Formats

Pattern	I	II	Format III	IV	V
P(XXX)	.415	.427	.492	.514	.304
P(YYY)	.152	.154	.206	.199	.065
P(BBB)	.046	.056	.016	.024	.247
P(XYY)	.026	.028	.023	.033	.007
P(YXX)	.026	.038	.032	.035	.011
P(BYY)	.021	.019	.017	.010	.016
P(BXX)	.034	.031	.018	.022	.054
P(XBB)	.026	.026	.012	.008	.053
P(YBB)	.018	.018	.002	.007	.027
P(XXY)	.021	.020	.024	.026	.009
P(XXB)	.027	.019	.023	.012	.032
P(YYX)	.014	.016	.020	.012	.008
P(YYB)	.015	.013	.010	.006	.008
P(BBY)	.011	.007	.002	.005	.013
P(BBX)	.013	.011	.006	.006	.034
P(XYX)	.019	.019	.022	.027	.008
P(XBX)	.028	.022	.015	.008	.026
P(YXY)	.012	.017	.011	.015	.007
P(YBY)	.014	.015	.014	.004	.012
P(BXB)	.011	.009	.006	.002	.023
P(BYB)	.010	.007	.002	.005	.013
P(XBY)	.006	.006	.006	.004	.002
P(XYB)	.009	.008	.007	.005	.007
P(YXB)	.007	.003	.006	.004	.001
P(YBX)	.007	.003	.005	.003	.003
P(BXY)	.004	.003	.003	.002	.006
P(BYX)	.005	.005	.001	.003	.002

Consistency

There should be more consistency from occasion two to three, than from one to two. Table 2 shows the proportion of total responses consistent from one occasion to the next. For each format more responses were changed from occasion one to two than from occasion two to three. Table 3 shows that in every case the probability of repeating on occasion two, a particular response from occasion

TABLE 2

Proportion of Responses Unchanged on Consecutive Pairs of Occasions

Format	First to Second	Second to Third
I	.714	.863
II	.723	.797
III	.799	.818
IV	.804	.852
V	.720	.784

one, ($P(X|X)$, $P(Y|Y)$, $P(B|B)$) was less than the probability of repeating a response on occasion three, from occasion two ($P(X|X)$, $P(Y|Y)$, $P(B|B)$).³ At the bottom of Table 3 it is shown that, for four of the five formats, the difference in sta-

TABLE 3

Sample Probabilities of Selected Consistency Patterns

Pattern	I	II	Format III	IV	V
$P(X X)$.802	.810	.864	.867	.770
$(PX -X)$.853	.875	.881	.903	.826
$P(X XX)$.896	.916	.913	.931	.881
$P(Y Y)$.683	.661	.771	.761	.570
$P(Y -Y)$.734	.747	.799	.807	.657
$P(Y YY)$.840	.842	.873	.917	.802
$P(B B)$.452	.500	.338	.393	.721
$P(B -B)$.533	.633	.385	.565	.784
$P(B BB)$.657	.757	.667	.686	.840
$P(X YX)$.578	.655	.653	.648	.579
$P(Y XY)$.481	.509	.442	.508	.318
$P(X BX)$.694	.721	.667	.846	.651
$P(Y BY)$.583	.613	.850	.556	.516
$P(B XB)$.433	.481	.364	.276	.654
$P(B YB)$.462	.500	.095	.500	.643
$P(X XY)$.352	.345	.423	.415	.364
$P(Y YX)$.267	.293	.224	.278	.368
$P(X XB)$.467	.407	.455	.400	.321
$P(Y YB)$.359	.417	.667	.286	.286
$P(X X)-P(Y Y)$.119	.149	.093	.106	.200
$P(X -X)-P(Y -Y)$.119	.128	.082	.096	.169

³ The following notation is used consistently throughout this paper:

- $P(X|X)$ —The probability of an X response on the *second* testing occasion, given an X response on *first* occasion.
- $P(X|X)$ —The probability of an X response on the *third* occasion given an X response on the *second* occasion.
- $P(X|X)$ —The probability of an X response on the *third* occasion given an X response on the *first* occasion.
- $P(X|XY)$ —The probability of an X response on the *third* occasion given an initial X response and a Y response on the *second* occasion.

bility between X (Majority) and Y (Minority) responses was less on occasion two ($P(X|X) - P(Y|Y)$) than on occasion one ($P(X|X) - P(Y|Y)$). Although all responses increased in stability, Y (less popular) responses tended to become relatively more stable on later occasions.

Differentiation of Individuals

Responses should be more equally distributed over the response alternatives on later occasions. In this study there were two estimates of response distribution: (a) the product of the endorsement proportions for the three responses for each item (i.e., $P(X) \times P(B) \times P(Y)$) and (b) the difference between the endorsement proportions of the X and Y responses for each item, (i.e., $P(X) - P(Y)$). Table 4 reports the mean values for these indices, for each format for each occasion. Increasing item variance would be shown by an increase in the index, $(P(X) \times P(B) \times P(Y))$ and by a de-

TABLE 4
Mean Values for Indices of Item Variances for Three Occasions

Format	Index	Occasion		
		1	2	3
I	$P(X) \times P(B) \times P(Y)$.0216	.0231	.0230
II	$P(X) \times P(B) \times P(Y)$.0182	.0186	.0187
III	$P(X) \times P(B) \times P(Y)$.0124	.0133	.0132
IV	$P(X) \times P(B) \times P(Y)$.0129	.0120	.0125
V	$P(X) \times P(B) \times P(Y)$.0187	.0185	.0188
I	$P(X) - P(Y)$.311	.288	.293
II	$P(X) - P(Y)$.299	.301	.300
III	$P(X) - P(Y)$.317	.323	.305
IV	$P(X) - P(Y)$.355	.333	.334

crease in the index $(P(X) - P(Y))$. When comparing the first and third testing occasion four of the five formats showed an increase on the first index and three of four formats⁴ showed a decrease on the second index. Increasing item variance thus seems to be demonstrated as a weak effect.

Analysis of Item Characteristics

The remainder of the analyses reported here represent attempts to extend and/or replicate previous studies of item parameters.

⁴ Format V was excluded since for this format the B responses were usually endorsed more often than the extreme responses.

Relationship between Endorsement Frequency and Consistency

Many investigators (e.g., Fiske, 1957b; Goldberg, 1963; Edwards and Walsh, 1963) have reported a strong relationship, for dichotomous items, between percent initial endorsement and percent consistent on retest. Usually this relationship is curvilinear—items with very high percent endorsement or very low percent endorsement are more stable on retest than are other items. This curvilinearity may be viewed as an artifact, however, since in the case of dichotomous scoring, low endorsement items are necessarily items for which the negative response alternative is most popular. In order to avoid this artifact, the relationship between endorsement and consistency was investigated for Majority (X) and Minority (Y) responses taken separately. Referring to Table 3, it can be seen that for each format, majority responses were more stable than minority responses. B responses were the least stable (with the exception of the "Continuum" format) and in the case of Format III and IV ("Uncertain" formats) the B responses were at about a random consistency level.

Omitting Format V, the correlations between majority endorsement, $P(X)$ and consistency, $P(X|X)$, were .48, .42, .43, and .54, for Formats I, II, III, and IV, respectively. For minority endorsement, $P(Y)$, the correlations were .41, .55, .08, and .01, respectively. Because of their similar patterns, Formats I and II were combined and Formats III and IV were combined. Figure 1 shows the relationship between endorsement and consistency for these two global formats ("In Between" (IB) and "Uncertain" (U)).

Three phenomena are apparent from Figure 1.

1. At all endorsement levels, X and Y responses to U-Format items were more stable than were X and Y responses to IB-Format items.
2. U-Format items exhibited a curvilinear relationship between endorsement and stability, while IB-Format items showed an essentially linear relationship.
3. At each comparable level, B responses were much less stable than were other responses, although IB-Format B responses showed the same general relationship between endorsement and consistency as did X and Y responses. The stability of U-Format B responses was essentially constant over the two

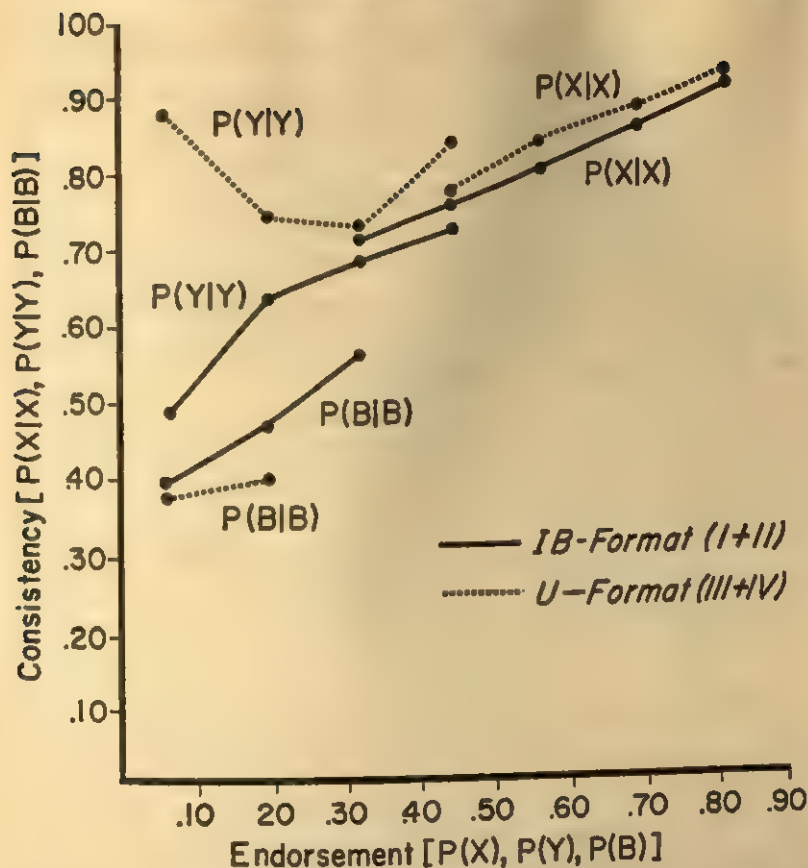


Fig. 1. Relationship between initial endorsement level of a response alternative and the probability of repeating the response on the second occasion.

endorsement levels represented and was at a random level.

The relationship between endorsement and consistency was, in general, the same for the second and third testing occasions as for the first and second. However, the regression lines are somewhat flatter and *the relationship for U-Format items becomes linear.*

Since $P(X)$ and $P(Y)$ are necessarily negatively related, the relationship between $P(X|X)$ and $P(Y|Y)$ would also tend to be negative. The actual correlations between $P(X|X)$ and $P(Y|Y)$ were $-.12$, $-.18$, $-.34$, and $-.00$ for Formats I, II, III, and IV, respectively. Given the pattern of the above results, it becomes apparent that it would be misleading to talk only about "item stability." Rather, we also should be concerned with response stability,

taking into account the format for presentation of response alternatives, as well as number of alternatives and item wording.

Continuum Format. With Format V items the B response tended to be more popular than with other formats. Figure 2 illustrates the relationship between initial endorsement and consistency for this format taking X, Y, and B responses separately. This again highlights the point that differences in format can influence initial endorsement and stability of response. While the B response in the continuum format is also generally an "In Between" response, these observed differences suggest that there is variation in individual and group reaction to the different formats.

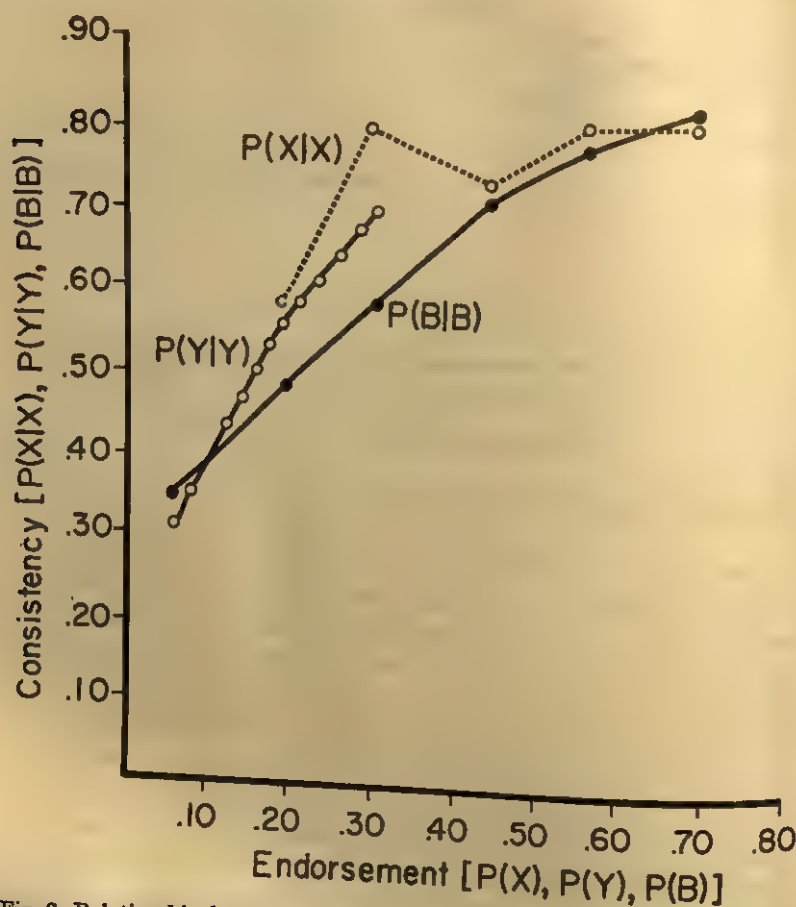


Fig. 2. Relationship between initial endorsement level and the probability of repeating a response on the second occasion for the Continuum Format (V).

The Meaning of B Responses. Edwards and Walsh (1963) have recently reported data which illustrate that "doubtful" responses are more often changed in a way that tends to increase the probability of a socially undesirable (less popular) response. Our results supported this finding. Considering items with at least two initial B responses which were changed on the second occasion, we compared $\Pr(X/(X+Y))$ with $\Pr(BX/(BX + BY))$. (The symbol BX represents those response patterns where the B response was initially endorsed and then changed to an X response on the second occasion.) The results are shown on Table 5. Figure 3 illustrates the same re-

TABLE 5

*Comparison of the Probability of the Popular Response
with the Probability of Changing from B to the Popular Response*

N	Format			
	I 81	II 18	III 10	IV 11
$\frac{X}{X+Y}$.682	.686	.665	.665
$\frac{BX}{BX+BY}$.568	.632	.544	.535
$\frac{B-X}{(B-X)+(B-Y)}$.583	.633	.470	.574
$\frac{-X}{(-X)+(-Y)}$.670	.690	.655	.663
$\frac{-BX}{(-BX)+(-BY)}$.602	.498	.566	.546

results over levels of $\Pr(X/(X+Y))$. It can be seen that B responses were more likely to change to Y than would be predicted on the basis of $\Pr(Y/(X+Y))$; i.e., *changed B responses give a closer to 50-50 "X - Y" split than the original "X - Y" split.*

Goldberg (1963) has proposed a model which states that there should be an inverse relationship between the number of B responses and the index $P(X) - P(Y)$. "(T:3a) Balanced items will be placed in a '?' category, or will be left blank, by more individuals than will extreme items" (Goldberg, 1963, p. 486). The closer the response endorsement split, the more the ambiguity (more difficulty in choosing between equal responses) of the item. Our results tended to support this hypothesis—the correlations were $-.18$ for IB-Format items, and $-.25$ for U-Format items. Figure 4 shows the

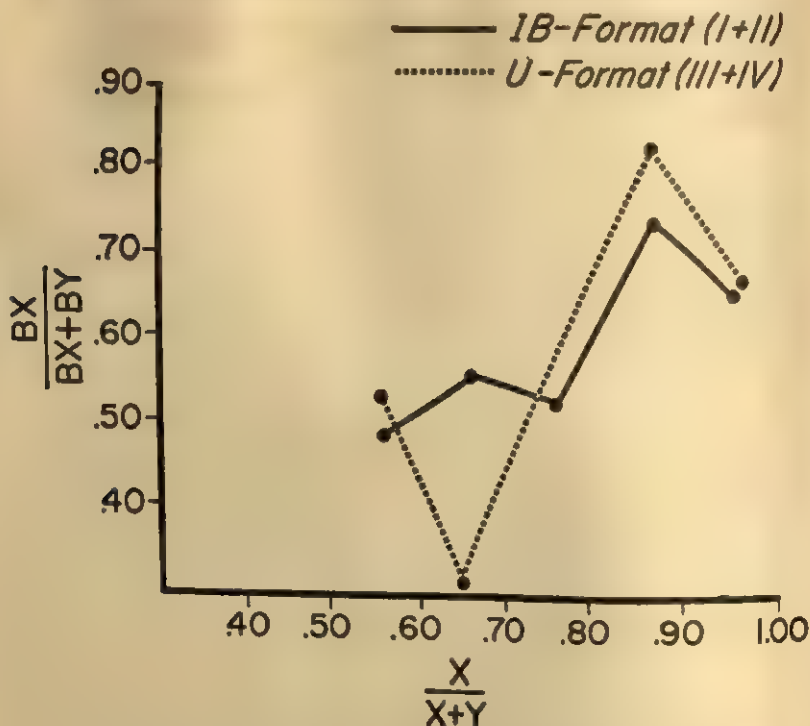


Fig. 3. Comparison of changed B response and initial X or Y responses.

relationship between $(P(X) - P(Y))$ and $P(B)$. Again, Figure 4 demonstrates that U-Format items elicited fewer B responses than did IB-Format items.

Summary and Conclusions

Eighty employment counselors completed the 16 PF, Form A, once a week on three consecutive weeks. Items were categorized into five response formats—based on wordings of response alternatives. Analyses were concerned with extensions of previous work on repeated testing effects, response formats, and response parameters.

Response Format Characteristics

The wording of the extreme response alternatives (X and Y) had no demonstrable effect on response characteristics. There is no evidence for deciding between an "Agree-Disagree" format and a "Choice" format.

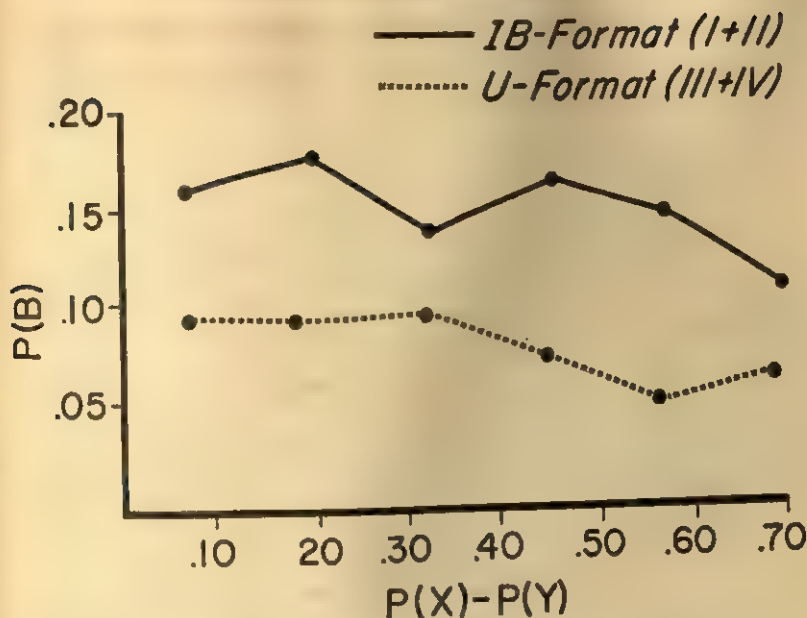


Fig. 4. Relationship between response split and proportion of B responses.

The Continuum Format provided a true three-choice situation with the middle response (B) covering a larger band of the trait dimension than was the case for the other formats. In terms of stability, the B responses showed the same trends as did the X and Y responses for this format. When the middle response was labelled "In-Between" (IB) it showed characteristics similar to the B response of the Continuum Format. On the other hand, when the middle response was labelled "Uncertain" (U) it showed essentially random stability. "Uncertain" responses showed re-test characteristics which were more typical of a "doubtful" or "?" response, i.e., low endorsement level and essential equality of $P(X|B)$, $P(B|B)$, and $P(Y|B)$. U-Format items showed more stability of X and Y responses, over all levels of initial endorsement, as well as higher endorsement of X and Y responses than did the other formats. In general, U-Format items showed characteristics of a true two-choice response situation.

Item Parameters

There was a moderate relationship between initial endorsement frequency and response stability for most categories of response. The

exception was in the case of minority responses in the U-Format—here there was a curvilinear relationship. Interpretation of this exception must await replication—especially since the relationship became linear when we compared endorsement frequency on the second occasion with stability of responses from the second to the third occasions. In general, the relationship between endorsement and stability becomes flatter on repeated testing.

Additional findings were as follows.

1. The closer the endorsement frequencies of X and Y responses, the more likely the endorsement of a B response.
2. B responses were, in general, less stable than X and Y responses.
3. On retest, a B response was more often changed to the minority response (Y) than would have been predicted from the relative initial endorsement frequency of the Y response.

Effects of Repeated Testing

The results of this study illustrate, on an item level, effects previously demonstrated through analyses of total scores and profiles (Howard, 1964). Responses became more stable and item variances tended to increase on repeated administrations of the same test. The conclusion is inevitable that responses to a later test administration yield a greater differentiation of individuals and more reliable individual differences than do responses to an initial test administration. The design of this study does not allow the elaboration of further explanation of these effects. Other studies, now in progress, attempt to establish the contribution of potential sources of increased "true score" variance with repeated test administrations. The validity of later test scores has yet to be established.

REFERENCES

- Buss, A. H. "The Effect of Item Style on Social Desirability and Frequency of Endorsement." *Journal of Consulting Psychology*, XXIII (1959), 510-513.
- Cattell, R. B., Saunders, D. R., and Stice, G. *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, Illinois: Institute for Personality and Ability Testing, 1957.
- Cronbach, L. J. "Response Sets and Test Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 475-494.
- Edwards, A. L. and Walsh, J. A. "Relationships between Various

- Psychometric Properties of Personality Items." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 227-238.
- Fiske, D. W. "The Constraints on Intra-Individual Variability in Tests." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVII (1957), 317-337 (a).
- Fiske, D. W. "An Intensive Study of Variability Scores." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVII (1957), 453-465. (b)
- Fiske, D. W. "The Inherent Variability of Behavior." In D. W. Fiske and S. R. Maddi (Eds.), *Functions of Varied Experience*. Homewood, Illinois: The Dorsey Press, 1961.
- Goldberg, L. R. "A Model of Item Ambiguity in Personality Assessment." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 467-492.
- Hanley, C. "Responses to the Wording of Personality Test Items." *Journal of Consulting Psychology*, XXIII (1959), 261-265.
- Howard, K. I. "Differentiation of Individuals as a Function of Repeated Testing." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIV (1964), 875-894.
- Wiener, D. N. "Subtle and Obvious Keys for the MMPI." *Journal of Consulting Psychology*, XII (1948), 164-170.
- Wiggins, J. S. "Strategic, Method, and Stylistic Variance in the MMPI." *Psychological Bulletin*, LIX (1962), 211-223.

INTERRELATIONSHIPS AMONG MMPI ITEM CHARACTERISTICS¹

JERRY S. WIGGINS

University of Illinois

AND

LEWIS R. GOLDBERG

University of Oregon and
Oregon Research Institute

THE present investigation is concerned with relationships among eight general categories of item characteristics which have been proposed to be important parameters in objective personality assessment. These categories are: (a) *group endorsement percentages* (Edwards, 1953; Gordon, 1953; Buss, 1959; Goldfried and McKenzie, 1962; Wiggins, 1962; Edwards and Walsh, 1963); (b) *rated social desirability values* (Heineman, 1952; Gordon, 1953; Edwards, 1957; Messick and Jackson, 1961) (c) *dispersion of desirability ratings* (Messick and Jackson, 1961; Edwards and Walsh, 1963); (d) *temporal stability* (Mittra and Fiske, 1956; Fiske, 1957; Edwards and Walsh, 1963; Goldberg, 1963; Goldberg and Rorer, 1963; Goldberg and Rust, 1963); (e) *item ambiguity* (Isard, 1956; Strong, 1962; Goldberg, 1963; Goldberg and Rorer, 1963; Gordon, 1953) (f) *direction of deviance* (Barnes, 1956a, 1956b; Berg, 1961; Sechrest and Jackson, 1962; Wiggins, 1962; Stricker, 1963); (g) *item*

¹This research was supported by Research Grant MH-07042-01 from the National Institute of Mental Health to the first author, and by Research Grant G-25123 from the National Science Foundation to the second author. Data analysis was carried out through the facilities of the Western Data Processing Center at the University of California at Los Angeles, the Digital Computer Laboratory at the University of Illinois, and the Statistical Laboratory at the University of Oregon. The authors wish to thank Richard Jones and Victor R. Lovell for their assistance on this project.

serial position (Cowen and Stiller, 1959; McGee and Komorita, 1963); and (h) *grammatical classifications* (Bergs and Martin, 1961; Brown and Adams, 1954; Buss and Durkee, 1957; Buss, 1959; Hanley, 1959; Elliott, 1961; Aiken, 1962; Goldfried and McKenzie, 1962; Stricker, 1963; Wiggins, 1963).

The distributions and interrelationships of these eight general categories were examined in the fixed pool of 566 items that comprises the MMPI. In addition to providing a more precise specification of the stimulus properties of the MMPI, such an investigation serves to highlight the virtues and limitations of the existing pool as a representative source of items for scale construction procedures.

Method

The item characteristics employed in the present study are listed in Table 1. Endorsement percentages (percent of the sample responding "true") were obtained from two college populations, a psychiatric group, and groups of subjects given instructions to answer the MMPI so as to make a desirable impression. Average social desirability ratings and their dispersions were taken from the published report of Messick and Jackson (1961). Stability values were based upon the percentage of college subjects whose response to the item remained the same on retesting after four weeks (Goldberg and Rorer, 1963). Item ambiguity values, expressed by the statistic Ambdex (Goldberg, 1963) which essentially corrects stability to take into account the item's endorsement percentage, were taken from Goldberg and Rorer (1963). Direction of deviance ("deviant true" vs "deviant false") was based on the least frequent response options of a preliminary sample of Minnesota normals (Hathaway and McKinley, 1951). Booklet number utilized the serial position of the item in the group form booklet. Relative temporal frequency was based on an extension of the work of Simpson (1944) who obtained judgments from high school and college students as to the percentage of time indicated by qualifying phrases (e.g., "hardly ever"); median ratings formed the basis of a scale ranging from "never" (00%) to "always" (99%). Negation was a dichotomous classification of positively vs. negatively phrased items. Sentence structure was a three-fold categorization, simple vs. compound vs. complex sentences. The number of words in the item was used as an index of sentence length. Voice was dichotomized into active vs.

TABLE 1

MMPI Item Characteristics Employed as Variables in the Present Study

Item Characteristic	Description
<i>Endorsement Percentages</i>	
Stanford men	105 college males
Stanford women	85 college females
Oregon men	95 college males
Oregon women	108 college females
Psychiatric men ^a	132 Palo Alto VA inpatients
Role-playing men	144 specially-instructed Stanford males
Role-playing women	106 specially-instructed Stanford females
<i>Desirability</i>	
Desirability values ^b	9-point ratings by 171 Penn. State students
Desirability dispersions ^b	Standard deviations of desirability ratings
<i>Stability</i>	
Stability: men	% of Oregon men giving the same response on 2nd administration
Stability: women	% of Oregon women giving the same response on 2nd administration
<i>Ambiguity</i>	
Ambdex: men	Item ambiguity values for Oregon men
Ambdex: women	Item ambiguity values for Oregon women
<i>Deviance and Order</i>	
Direction of deviances ^c	Least frequent response option (T or F) in Minnesota normals
Booklet number	Item number in group form booklet
<i>Grammatical Classification^d</i>	
Relative temporal frequency	% of time indicated by qualifying phrases
Negation	Positive vs. negative declarative forms
Sentence structure	Simple vs. compound vs. complex sentences
Sentence length	Number of words in the item
Voice	Active vs. passive voice
Tense	Future, present, present perfect, past
Person	First person vs. other

^aFrom Ullmann, 1962.^bFrom Messick and Jackson, 1961.^cFrom Hathaway and McKinley, 1951.^dGrammatical classifications were prepared by Victor R. Lovell and Michael Benjamin.

passive. Four categories of tense were employed, past vs. past perfect vs. present vs. future. Person was a dichotomy of first vs. other. A more detailed description of the item categories employed in the present study is available elsewhere (Wiggins, 1963).

Results

Before calculating correlation coefficients, the single and joint frequency distributions of the 22 variables listed in Table 1 were examined. Of the dichotomous variables, three (voice, tense, person) had such disproportionately large frequencies for one category that

calculation of biserial correlations was not appropriate. Since an overwhelming majority of MMPI items are written in the active voice of the first person and in the present tense, these three variables were eliminated from subsequent analyses.

Many of the continuous variables were skewed in their distributions and in some cases their regression on other variables in the study was clearly not linear. For this reason, three separate correlation coefficients were computed among all variables: (a) Pearson product-moment coefficients for continuous data; (b) Pearson product-moment coefficients for grouped data; and (c) correlation ratios (eta) for grouped data. Computation of the latter two correlations permitted an F test for curvilinearity (McNemar, 1955, pp. 268-275). Table 2 presents the means, standard deviations, and inter-correlations among the 19 variables which were analyzed. Significant product-moment correlations ($r > \pm .138$; $p < .001$) appear in the upper right of the matrix. Correlation ratios (etas) which significantly departed ($p < .001$) from their corresponding product-moment correlations are given in the lower left of the matrix. However, it should be borne in mind that the lack of significant departure of eta from r is no absolute guarantee of linearity in the bivariate distribution (McNemar, 1955, p. 275).

The magnitude of the standard deviations associated with the mean values in Table 2 indicates that the wide range of values for these variables is poorly summarized by their arithmetical average. Nevertheless, it is of some interest to characterize the item properties of the "typical" MMPI item. The present data suggest that the "average" MMPI item is written in the first person singular, present tense. It is a positive declarative sentence in active voice implying that the action occurs about 50 percent of the time. The item is written as a compound sentence containing 11 words. Such an item will be answered "true" by 42 percent of college students and 46 percent of psychiatric patients. The direction of deviant response for both groups would consequently be "true." Upon repeated administrations of the item, 87 percent of college students will answer it with the same response as they did before. The item appears to be relatively unambiguous and is rated as being "mildly undesirable."

Perhaps of more interest is the manner in which the item characteristics relate to each other within the collection of 566 items that

TABLE 2
Intercorrelations among 19 MMPI Item Characteristics^a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Stanford men	1	—	91	97	89	84	90	83	80	18	-20	-17		-78		-16	25		
Stanford women	2		—	91	97	82	87	85	81	17	-16	-18		-73		-15	20		
Oregon men	3			—	91	85	91	83	82	18	-22	-18		-77		-16	22		
Oregon women	4				—	80	86	85	80	16	-16	-17		-73			17		
Psychiatric men	5					—	78	72	70		-26	-22		-67			20		
Role-playing men	6						96	92	24	24	-18	-16		-79			21		
Role-playing women	7						—	92	24	-15	-16			-76			21		
Desirability values	8							—	19			-18		-72			17		
Desirability dispersions	9						24		—					-22		-28	-21		
Stability: men	10	70	64	75	64	59	58	55	48			75	-26		-16			-16	-15
Stability: women	11	59	71	62	74	54	55	57	48						-17			-18	-31
Ambdex: men	12	24	24	23	21	20	25			48	35	—	37					15	15
Ambdex: women	13		23		22					39	54	—						14	14
Direction of deviance	14																	-30	14
Booklet number	15			20			21	24	36										
Relative temporal freq.	16	30	27	28	26		31	34	31								-23		
Negation	17																		
Sentence structure	18		28																61
Sentence length	19	24					20	26	30	25									
Mean		.40	.42	.42	.42	.46	.43	.43	4.64	1.54	.87	.87	.56	.57	.63	283.50	.49	1.14	1.91
Standard Deviation		.30	.30	.31	.31	.23	.31	.34	1.53	.33	.07	.08	.21	.26	.48	163.39	.19	.35	.99

^aAbove the diagonal are listed product-moment correlations significantly different from zero ($p < .001$). Below the diagonal are listed etas significantly different from their corresponding linear r 's ($p < .001$).

constitute the MMPI item pool. These results are summarized in the main body of Table 2; each of the general classes of item characteristics included in the present study will be discussed in turn.

Endorsement Percentages. Examination of the means and standard deviations for endorsement percentages reveals a striking comparability for different populations. The means and standard deviations for the four college groups are virtually identical. Role-playing instructions tend to increase "true" answers very slightly. The endorsement percentages of the psychiatric group differ from the college groups by, at most, five percentage points; however, the variance of the endorsement percentages for psychiatric patients is less than that of the several college groups.

Correlations among endorsement percentages for the several groups permit comparison of the relative importance of sex, regional location, and psychological health. Within the two college populations represented, endorsement percentages by the same sex from different institutions are more highly correlated (.97) than those by different sexes attending the same institution (.91). Endorsement percentages across both sex and institution are still highly similar (.89 and .91). The endorsement percentages of male psychiatric patients are correlated in the .80's with those of college students of both sexes. The magnitude of the correlation between the endorsement percentages of such diverse groups as Stanford coeds and Palo Alto VA psychiatric men has rather serious implications regarding the probable success of empirical scales derived by the method of contrasted groups (Ullmann and Wiggins, 1962). The correlation between men and women role-players from the same institution is high (.96). Correlations between role-playing and standard instruction endorsement percentages tend to be about the same when the role-playing and standard groups are of the same sex but from different institutions (.91, .85) as when the role-playing and standard groups are of different sexes but from the same institution (.83, .87).

Desirability Ratings. The correlations between endorsement percentages and desirability ratings tend to be highest in role-playing groups (.92), next highest in college samples (e.g., .81), and least high in a psychiatric sample (.70). Role-playing groups tend to endorse items on which there is the greatest dispersion of opinion as to their desirability values (.24). This tendency is present to a lesser

degree in standard instruction groups. Among psychiatric men, the same trend prevails, except that the very highest dispersion items are less frequently endorsed; the relationship is a complex one, not easily described by the eta of .24. Dispersion of opinion as to item desirability is greatest when the item is stated in negative form (.26). Desirability ratings are curvilinearly related to their own dispersions; both high and low desirability items have greater dispersions than medium desirability items ($\eta = .33$). Undesirable items tend to have slightly more complex sentence structure ($-.15$).

Stability and Ambiguity. The relationships between item stability and endorsement percentages for the various samples reveal the most striking departures from linearity of any relationships in this study. In accord with theoretical expectations and previous empirical findings (Goldberg, 1963), items of extreme endorsement percentages tend to be the most stable. The highest etas are found between item stability and endorsement percentages for the same sample (.75 for men, .74 for women). The curvilinear relationship between stability and endorsement percentage declines slightly between different college samples (.70 and .71), decreases across sex but within college populations (e.g., .62) and decreases again slightly for role-playing and psychiatric groups (e.g., .58). The curvilinear relationship between item stability and rated desirability is substantial (.48).

As Table 2 reveals, there is a slight curvilinear relationship between endorsement percentages and Ambdex values; extreme items have slightly greater Ambdex values (e.g., $\eta = .22$). Note that these relationships are much less than those between endorsement percentages and item stability. Although Ambdex was constructed so as to be independent of endorsement percentage, this independence does not appear to be completely realized over the entire 566 item MMPI pool, at least for samples of this size. Table 3 presents the interrelationships between endorsement percentages, stability values, and ambiguity values for the Oregon samples, for sub-pools of progressively less extreme items. As Table 3 indicates, when the most extreme MMPI items are progressively eliminated, the relationship between Ambdex and stability increases while the relationship between endorsement percentage and Ambdex, never great to begin with, remains about the same. The strong curvilinear relationship between endorsement and stability decreases as the item

TABLE 3

The Relationship between Endorsement Percentage (E), Stability (S), and Ambdex (A), for the Oregon Samples, as a Function of Different Item Pools Progressively Decreasing in Extremeness

	Males (M)			Females (F)			M vs F	Endorsement %	# Items
	E	S	A	E	S	A			
E	—	-.22 ^a	-.01	—	-.17	-.09	91 ^a		
S	.75 ^b	—	-.26	.74	—	-.17	.75	0-100	566
A	.23	.48	—	.22	.55	—	.37		
E	—	-.19	.09	—	-.13	.03	.85		
S	.59	—	-.69	.59	—	-.68	.65	5-94	433
A	.19	.69	—	.14	.69	—	.52		
E	—	-.15	.13	—	-.12	.01	.80		
S	.42	—	-.82	.46	—	-.82	.58	10-89	359
A	.25	.80	—	.16	.82	—	.57		
E	—	-.14	.07	—	-.06	-.05	.76		
S	.31	—	-.89	.37	—	-.90	.59	15-84	307
A	.21	.87	—	.18	.89	—	.61		
E	—	-.12	.06	—	-.03	-.06	.72		
S	.22	—	-.95	.28	—	-.94	.59	20-79	243
A	.19	.91	—	.19	.92	—	.62		
E	—	-.08	.03	—	.04	-.09	.58		
S	.19	—	-.98	.31	—	-.98	.62	25-74	187
A	.11	.94	—	.20	.95	—	.63		
E	—	-.10	.10	—	.01	-.06	.51		
S	.28	—	-.99	.14	—	-.99	.61	30-69	140
A	.25	.95	—	.11	.96	—	.61		

^aAbove the diagonal are listed r 's.

^bBelow the diagonal are listed s 's.

^cCorrelations for M vs F are linear r 's.

pool becomes more balanced. As progressively more extreme items are eliminated from the MMPI pool, the correlation between male and female endorsement percentages (as well as between male and female stability values) goes down, while the correlation between male and female Ambdex values rises dramatically. Interrelationships among the other item characteristics utilized in this study do not display any noticeable change when extreme items are eliminated.

Stability is linearly related to sentence length in both male (— .31) and female (— .28) groups; shorter sentences tend to be answered more consistently. Items with complex sentence structures tend to be less stable over time (— .18). Items with more complex sentence structure, understandably, are of greater length (.61). Items appearing at the beginning of the test booklet are more stable than those appearing at the end (— .16, — .17). Ambiguous items

tend to be longer (.15, .14). Items are ambiguous for women, though not for men, when the items are undesirable (— .18); items are ambiguous for men, though not for women, when the items are complex in sentence structure (.15).

Deviance and Order. Direction of deviance (true or false) is highly, but not perfectly, related to desirability ratings (— .72). Deviance is related to endorsement percentages in the various groups in much the same manner as desirability is related to endorsement in the same groups. Desirability ratings of deviant true items have less dispersion than ratings of deviant false items (— .22).

Items appearing early in the group form booklet tend to have more extreme desirability ratings ($\eta^2 = .24$) and to have slightly smaller desirability dispersions (— .27) than items appearing toward the end of the booklet. Oregon women tend to have more extreme endorsement percentages for early booklet items ($\eta^2 = .20$); role-playing women, on the other hand, tend to endorse less frequently both items at the beginning and a later block in the middle of the booklet, as compared to all other items ($\eta^2 = .21$).

Grammatical Characteristics. Many of the scattered relationships between grammatical and other item characteristics have already been mentioned. In addition, there are several relationships which involve the negative versus positive item phrasing variable and the variable of relative temporal frequency. Items stated in the negative form tend to be deviant when answered false (— .30). There is also a slight tendency for items with negative phrasing to be judged more desirable (.17). A slight, but consistent, positive correlation exists between negative item phrasing and endorsement percentage in all groups (.17 to .25). Items which are negatively stated tend to imply less frequent occurrence of the behavior in question (— .23).

One of the most interesting nonlinear relationships emerging from this study is that between the relative temporal frequency implied by an item and its endorsement percentage. The complex association between these two variables may be seen in Table 4 which presents average endorsement percentages for the seven subject groups on five representative temporal classifications, along with the corresponding average desirability value for each classification. Reading across the rows of Table 4, it can be seen that endorsement

TABLE 4
Average Endorsement Percentages and Desirability Values for Item Groups Differing in Relative Temporal Frequency

Subjects	Never (0%) n = 22		At Times— Sometimes (20-21%) n = 47		Unqualified (50%) n = 391		Often (78%) n = 30		Most of the Time—Always (87-99%) n = 18	
	\bar{x}	σ	\bar{x}	σ	\bar{x}	σ	\bar{x}	σ	\bar{x}	σ
Stanford men	.66	(.32)	.34	(.28)	.41	(.30)	.22	(.15)	.35	(.34)
Stanford women	.66	(.32)	.39	(.29)	.41	(.30)	.25	(.17)	.38	(.36)
Oregon men	.66	(.33)	.38	(.29)	.42	(.31)	.24	(.18)	.37	(.35)
Oregon women	.66	(.31)	.38	(.30)	.42	(.30)	.27	(.16)	.36	(.36)
Psychiatric men	.62	(.23)	.42	(.24)	.47	(.23)	.35	(.17)	.38	(.22)
Role-playing men	.70	(.27)	.30	(.24)	.44	(.31)	.27	(.18)	.41	(.39)
Role-playing women	.73	(.25)	.27	(.25)	.45	(.34)	.25	(.20)	.42	(.43)
Desirability values	5.99	(1.08)	3.93	(1.16)	4.69	(1.57)	4.09	(.73)	4.85	(1.99)

is highest for the least frequent qualifier ("Never"), falls off rapidly at "Sometimes," rises for items that are unqualified with respect to frequency, falls dramatically at "often" and increases slightly at "always."

Although the foregoing row trends hold for all subject groups, inspection of the columns reveals interesting inter-group differences for several temporal classifications. By considering the overall level of endorsement along with observed differences among college, psychiatric, and role-playing groups, additional characterizations of the temporal classifications may be made. Items qualified by "never" tend to be endorsed by the majority of all groups, ranging from psychiatric patients (.62) to college students (.66) to role-players (.72). This ordering of subject groups would suggest that such items are perceived as "desirable," and this is clearly borne out by the high desirability mean value given in the last row (5.99). Items qualified by "sometimes" are endorsed by the minority of all groups, ranging from role-players (.29) to college students (.37) to psychiatric patients (.42). This ordering of subject groups suggests that such items are perceived as "undesirable," and this is supported by the low desirability mean (3.93). Items which are unqualified as to temporal frequency (the vast majority of MMPI items) are closest to 50 percent endorsement, ranging from college students (.42) to role-players (.45) to psychiatric patients (.47). These differences between groups are neither large nor readily interpretable in terms of direction of desirability; the items are on the average neutral in desirability (4.69). Items qualified by "often" are the least popular of all, ranging from college students (.24) to role-players (.26) to psychiatric patients (.35). Considering the generally low endorsement percentages and the slight differences between college and role-playing groups, such items appear to be both undesirable (4.09) and obvious. Items qualified by "always" are more popular overall and show only slight differences between college students (.37), psychiatric patients (.38), and role-players (.41). Such items tend to be of neutral (4.85) desirability value.

The standard deviations of the group endorsement percentages are largest for the "always" classification and smallest for the "often" classification. As was found in the main analysis, variability of endorsements tends to be greatest in college groups and smaller in role-playing and psychiatric groups, respectively. Ex-

ceptions found in the present analysis are the uniformly low standard deviations found in the "often" category and the relatively large standard deviations of the role-playing groups found in the "always" category.

For some reason more complex sentences are slightly less frequently endorsed by Stanford women, especially role-playing women ($-.16$); the relationship is not apparent for Oregon women or any of the male groups. Shorter items tend to have slightly higher desirability ratings, to be more often deviant when answered false, and to be endorsed slightly more often, but the relationships are not completely linear and are difficult to describe.

Discussion

The present findings tend to substantiate several previously reported relationships among classes of item characteristics, to qualify some previous findings, and to reveal many trends among item variables never previously studied. The striking comparability of item endorsement percentages from diverse groups under different instructional sets is a two-edged sword. The similarity among different college groups suggests that inter-institutional comparisons among college students on the MMPI may be made with some confidence (Goodstein, 1954). However, the similarity among *all* groups studied suggests that the MMPI item pool may be inefficient as a source of group discriminative items (Ullmann and Wiggins, 1962). The social desirability variable, as measured by group ratings, is substantially related to endorsement in the various groups in a manner consistent with the formulations of Edwards (1957). The relation of desirability to the grammatical characteristics of negative phrasing and sentence structure is consistent with previous reports of the sensitivity of this variable to changes in item structure or tone (Buss, 1959; Hanley, 1959). The relations found between the dispersion of social desirability ratings on the one hand and endorsement percentages and the social desirability ratings themselves on the other hand, underscore the importance of considering dispersions in conjunction with mean rated values (Messick and Jackson, 1961; Edwards and Walsh, 1963).

Edwards and Walsh (1963) analyzed some relationships among six item parameters for 176 miscellaneous personality inventory

items administered twice to 110 male and 111 female paid college students. The same 176 items were also rated for their social desirability on two occasions by a similar but separate group of 47 males and 48 females. Since four of the six item parameters computed by Edwards and Walsh are included in the present study (endorsement percentages, social desirability mean ratings, dispersions of the desirability ratings, and item stability), some findings from both studies can be compared.

The major point of congruence between the two studies stems from the already noted tendency of different samples to show similar endorsement percentages for any one item pool. Edwards and Walsh reported correlations in endorsement percentages between male and female samples of .96 (first administration) and .97 (second administration); the correlation from the present study between male and female role-players (.96) is identical to theirs, and the male vs. female correlations from standard administrations of the MMPI to Stanford and Oregon samples are quite similar (.91).

However, some rather striking differences emerged from the two studies, emphasizing the extent to which the relationships among item properties are a function of the particular pool of items investigated. The most striking such discrepancy concerns the relationship between the dispersion of desirability ratings and item stability. That this relationship should be positive and substantial has been postulated from two rather different theoretical models (Edwards and Walsh, 1963; Goldberg, 1963). However, while the predicted relationship was confirmed by Edwards and Walsh (.46), it did not hold up in the present study (.05 and .04, for males and females respectively).

Moreover, Edwards and Walsh found essentially no linear relationship between item stability on the one hand and endorsement percentages ($-.01$) and desirability ratings ($-.03$) on the other. It seems likely that the slight, but statistically significant, linear correlations found in the MMPI (e.g., $-.22$ for Oregon males and $-.17$ for Oregon females) are simply a function of the peculiarities of this pool of items, specifically of the fact that the MMPI item pool is not equally balanced on direction of deviance. When direction of deviance is balanced, the linear correlation should approach zero (Goldberg, 1963). The curvilinear relationships between desirability ratings and stability are similar (.48 in the present study vs. .62 in

Edwards and Walsh) and in accord with previous findings (Goldberg, 1963).

The characteristic of direction of deviance or non-communality of response has been emphasized by those subscribing to the Deviation Hypothesis (Barnes, 1956a, 1956b; Berg, 1961). The over-representation of deviant true items within the MMPI pool has been previously noted (Wiggins, 1962). Because of its overlap with both desirability and endorsement, Wiggins (1962) has characterized this variable as a "gross" measure. The present findings are compatible with the interpretation of deviance as a similar but not equivalent measure to both desirability and endorsement. Partial support was found in the present study for an order effect operating with desirability ratings and dispersions (Cowen and Stiller, 1959) as well as a slight confounding between order and endorsement in female groups.

The present study provides some additional support for the view that grammatical characteristics enter into relations with many other item characteristics (Brown and Adams, 1954; Buss and Durkee, 1957; Buss, 1959; Hanley, 1959; Elliott, 1961; Stricker, 1963). Since many of the grammatical characteristics employed in the present study were quite gross indices, it would seem likely that more significant relations might emerge were more refined scaling procedures applied to grammatical characteristics.

In addition to its more general findings regarding relationships among item characteristics, the present study has implications for the MMPI as a source of items for various assessment purposes. Ideally, an item pool should be assembled in such a way as to systematically sample both the substantive domains of interest and the many possible "facets" of item characteristics which may interact with substantive domains (Loevinger, 1957). From considerations of both construct and predictive validity, it is becoming apparent that the MMPI item pool was not assembled in a maximally efficient manner. Over- and under-representation of certain classes of desirability, endorsement, ambiguity, and grammatical characteristics tends to make the item pool unnecessarily homogeneous and may, in part, contribute to rather severe restrictions in criterion group discriminations. The fortuitous confounding of such item characteristics with substantive dimensions (Block, 1962; Wiggins, 1962) has created interpretative problems (Edwards, 1957; Jack-

son and Messick, 1961) which may never be satisfactorily resolved within any fixed item pool. It is of some importance to note that such shortcomings are not peculiar to the MMPI but would, no doubt, become apparent in any other personality inventory which was subjected to the scrutiny which the MMPI has enjoyed in the last decade. It is hoped that studies such as the present one will stimulate research at the item level, specifically attempts to experimentally manipulate item parameters through item rephrasing. It may well be that only through such experimental study can enough solid evidence be accumulated to allow future personality inventory constructors to sample all item characteristics related to the logical and predictive validity of such instruments.

Summary

This study attempted to provide a more precise specification of the stimulus properties of the MMPI and to highlight the virtues and limitations of the existing pool as a representative source of items for scale construction procedures. To this end, means, standard deviations, and intercorrelations were computed among 22 variables representing eight general categories of item characteristics. The results tended to substantiate several previously reported relationships among classes of item characteristics, to qualify some previous findings, and to reveal many trends among variables not previously noted. More specifically, it was suggested that the over- and underrepresentation of certain item characteristics found in the MMPI pool tends to make the pool unnecessarily homogeneous and that this may contribute to rather severe restrictions on possible criterion group discriminations.

REFERENCES

- Aiken, L. R., Jr. "Frequency and Intensity as Psychometric Response Variables." *Psychological Reports*, XI (1962), 535-538.
- Barnes, E. H. "Factors, Response Bias, and the MMPI." *Journal of Consulting Psychology*, XX (1956), 419-421. (a)
- Barnes, E. H. "Response Bias in the MMPI." *Journal of Consulting Psychology*, XX (1956), 371-374. (b)
- Berg, I. A. "Measuring Deviant Behavior by Means of Deviant Response Sets." In I. A. Berg and B. M. Bass (Editors), *Conformity and Deviation*. New York: Harper, 1961.
- Bergs, L. P. and Martin, B. "The Effect of Instructional Time Interval and Social Desirability on the Validity of a Forced-

- Choice Anxiety Scale." *Journal of Consulting Psychology*, XXV (1961), 528-532.
- Block, J. "Unconfounding Meaning, Acquiescence and Social Desirability in the MMPI." Unpublished manuscript, University of California, Berkeley, 1962.
- Brown, D. R. and Adams, J. "Word Frequency and the Measurement of Value Areas." *Journal of Abnormal and Social Psychology*, IL (1954), 427-430.
- Buss, A. H. "The Effect of Item Style on Social Desirability and Frequency of Endorsement." *Journal of Consulting Psychology*, XXIII (1959), 510-513.
- Buss, A. H. and Durkee, A. "An Inventory for Assessing Different Kinds of Hostility." *Journal of Consulting Psychology*, XXI (1957), 343-349.
- Cowen, E. L. and Stiller, A. "The Social Desirability of Trait Descriptive Terms: Order and Context Effects." *Canadian Journal of Psychology*, XIII (1959), 193-199.
- Edwards, A. L. "The Relationship between the Judged Desirability of a Trait and the Probability that the Trait Will Be Endorsed." *Journal of Applied Psychology*, XXXVII (1953), 90-93.
- Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden Press, 1957.
- Edwards, A. L. and Walsh, J. A. "Relationships between Various Psychometric Properties of Personality Items." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 227-238.
- Elliott, Lois L. "Effects of Item Construction and Respondent Aptitude on Response Acquiescence." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXI (1961), 405-415.
- Fiske, D. W. "The Constraints on Intra-Individual Variability in Test Responses." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVII (1957), 318-337.
- Goldberg, L. R. "A Model of Item Ambiguity in Personality Assessment." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 467-492.
- Goldberg, L. R. and Rorer, L. G. "Test-Retest Item Statistics for Original and Reversed MMPI Items." *Oregon Research Institute Research Monograph*, III, No. 1 (1963), Eugene, Oregon.
- Goldberg, L. R. and Rust, R. M. "Intra-Individual Variability in the MMPI-CPI Common Item Pool." *Oregon Research Institute Research Bulletin*, III, No. 3 (1963), Eugene, Oregon.
- Goldfried, M. R. and McKenzie, J. D., Jr. "Sex Differences in the Effect of Item Style on Social Desirability and Frequency of Endorsement." *Journal of Consulting Psychology*, XXVI (1962), 126-128.
- Goodstein, L. D. "Regional Differences in MMPI Responses among Male College Students." *Journal of Consulting Psychology*, XVIII (1954), 437-441.
- Gordon, L. V. "Some Interrelationships among Personality Item Characteristics." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XIII (1953), 264-272.

- Hanley, C. "Responses to the Wording of Personality Test Items." *Journal of Consulting Psychology*, XXIII (1959), 261-265.
- Hathaway, S. R. and McKinley, J. C. *The Minnesota Multiphasic Personality Inventory Manual* (Revised Edition). New York: Psychological Corporation, 1951.
- Heineman, C. E. Unpublished Materials. Iowa City: State University of Iowa, 1952.
- Isard, E. S. "The Relationship between Item Ambiguity and Discriminating Power in a Forced-Choice Scale." *Journal of Applied Psychology*, XXXX (1956), 266-268.
- Jackson, D. N. and Messick, S. "Acquiescence and Desirability as Response Determinants on the MMPI." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 771-790.
- Loevinger, Jane. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports*, III (1957), 635-694.
- McGee, R. K. and Komorita, S. S. "The Influence of Intra-Test Variables on Acquiescent Responses." Paper presented at Southeastern Psychological Association meetings, 1963.
- McNemar, Q. *Psychological Statistics* (Second Edition). New York: Wiley, 1955.
- Messick, S. and Jackson, D. N. "Desirability Scale Values and Dispersions for MMPI Items." *Psychological Reports*, VIII (1961), 409-414.
- Mitra, S. K. and Fiske, D. W. "Intra-Individual Variability as Related to Test Score and Item." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVI (1956), 3-12.
- Sechrest, L. and Jackson, D. N. "The Generality of Deviant Response Tendencies." *Journal of Consulting Psychology*, XXVI (1962), 395-401.
- Simpson, R. "The Specific Meanings of Certain Items Indicating Differing Degrees of Frequency." *Quarterly Journal of Speech*, XXX (1944), 328-330.
- Stricker, L. J. "Acquiescence and Social Desirability Response Styles, Item Characteristics, and Conformity." *Psychological Reports*, XII (1963), 319-341.
- Strong, E. K., Jr. "Good and Poor Interest Test Items." *Journal of Applied Psychology*, XLVI (1962), 269-275.
- Ullmann, L. P. "An Empirically Derived MMPI Scale which Measures Facilitation-Inhibition of Recognition of Threatening Stimuli." *Journal of Clinical Psychology*, XVIII (1962), 127-132.
- Ullmann, L. P. and Wiggins, J. S. "Endorsement Frequency and the Number of Differentiating MMPI Items to Be Expected by Chance." *Newsletter for Research in Psychology*, IV (1962), 29-35.
- Wiggins, J. S. "Strategic, Method, and Stylistic Variance in the MMPI." *Psychological Bulletin*, LIX (1962), 224-242.
- Wiggins, J. S. *Manual for the MMPI Item Characteristic Deck*. Technical Report No. 1, PHS Grant MH 07042-01. Urbana: University of Illinois, July, 1963.

COMMUNALITY AND FAVORABILITY AS SOURCES OF METHOD VARIANCE IN THE MMPI¹

JERRY S. WIGGINS

University of Illinois

and

VICTOR R. LOVELL

Stanford University

THE research to be reported has been guided by a taxonomy for components of test variance which differentiates among several hypothetical sources of influence thought to contribute to scores on objective personality tests (Wiggins, 1962a, 1962b, 1962c). Briefly, a distinction is made among sources of variance arising from: (a) the *strategy* of scale construction (e.g., use of contrasted criterion groups); (b) *stylistic* response tendencies on the part of subjects (e.g., bias to answer "true"); (c) item *content* (operationally defined by scaling judges' ratings of substantive dimensions); and (d) *method* variance arising from the idiosyncratic distribution of item parameters (e.g., endorsement frequencies) that characterizes a given item pool.

The present study reports an attempt to analyze two sources of method variance in the MMPI. Specifically, the item parameters of *communality* (endorsement frequency) and rated *favorability* formed the basis of scale construction procedures designed to il-

¹ This investigation was supported in part by a research grant, MH 07042-01, from the National Institute of Mental Health of the National Institutes of Health, Public Health Service.

We are indebted to Wesley C. Becker, Lewis R. Goldberg, and Leonard G. Rorer for their helpful criticisms of an earlier draft of this paper.

illuminate the role of these parameters as determinants of the structural characteristics of the MMPI.

Method

To insure some generality of findings, the item characteristics under investigation were determined from two rather different samples and applied to still a third. Goldberg and Rorer (1963) computed item endorsement percentages separately in groups of 95 male and 108 female undergraduates at the University of Oregon. For each of the 566 items of the MMPI, this is expressed simply as the proportion of subjects from the total group who answered "true." Messick and Jackson (1961b) obtained desirability ratings for each MMPI item from 171 Pennsylvania State University undergraduate men and women. These nine-point ratings, ranging from "extremely undesirable" to "extremely desirable" were then transformed to equal interval scale values by the method of successive intervals. In the present study, endorsement percentages of Oregon students and desirability ratings of Pennsylvania State students were used to construct special scales which were, in turn, scored from the MMPI protocols of 250 Stanford University undergraduate men.

Scale construction was facilitated by the availability of an MMPI Item Characteristic Deck (Wiggins, 1964a) which consists of 566 IBM cards each containing coded information on item characteristics for the corresponding 566 items of the MMPI. Using 20 intervals of five for the Oregon endorsement percentages and 16 intervals of one-half for the Penn State desirability values, cross-frequency tabulations were made on an IBM counter-sorter. These tabulations yielded a bivariate distribution of MMPI items on endorsement percentage by social desirability.

Inspection of the scatter plot of endorsement by desirability suggested that it was markedly linear in form, as has been emphasized by Edwards (1953). Nevertheless, departures from this linear association of endorsement and desirability were frequent enough to justify the investigation of item pools representing diverse combinations of endorsement and desirability values. The scatter plot was divided in such a way as to create endorsement-desirability categories which would contain sufficient numbers of items to comprise scales, and which would pose different choices in terms

of responding consistently to item communality and item desirability. Figure 1 illustrates the manner in which the scatter plot was sectioned to define 13 non-overlapping endorsement-desirability categories.

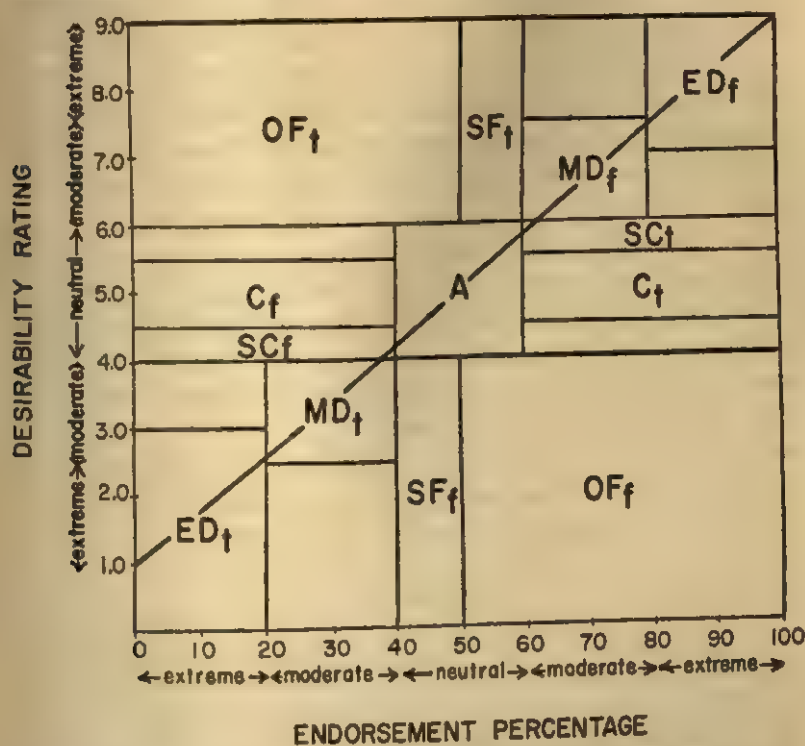


Figure 1. Partitioning of the scatter plot of endorsement percentage by desirability rating.

Although endorsement percentages and desirability values do not share a common metric, it is common to characterize intervals within each of the scales by similar designations. That is, social desirability values from 4.0 to 6.0 and endorsement percentages from 40 to 60 are thought of as "neutral" in value (Wiggins, 1962b). Similarly arbitrary intervals of moderate and extreme low and moderate and extreme high are indicated on both coordinates of Figure 1. The dotted diagonal line in Figure 1 passes from the lower left corner (low endorsement, low desirability) through the point of neutral endorsement (50 percent), neutral desirability (5.0) up to the upper right of the plot (high endorsement, high de-

sirability). By reference to this diagonal, it is possible to consider three combinations of endorsement and desirability values; endorsement may "equal" (fall on the diagonal), be more extreme than, or be less extreme than the corresponding desirability values.

The 13 endorsement-desirability categories chosen by this method are described more fully in Table 1. Where endorsement and desirability values are equal and extreme, the item pools are characterized by *deviance*. Where desirability values are more extreme than endorsements, the item pools are characterized by *favorability*. Where endorsements are more extreme than desirability values, the pools have been indicated as reflecting *communality*. In devising scales from these pools, items were keyed in such a way as to represent both directions of each characteristic. Thus, for "Extreme Deviance" there is a pool of extremely unfavorable, unpopular items keyed true (ED_t) and a pool of extremely favorable, popular items keyed false (ED_f). The same is true for each category of Favorability and Communality. The final category (A) contains neutral and approximately equal values of endorsement and desirability and has been keyed true according to the usual definition of an "acquiescence" scale (Wiggins, 1962b).

TABLE 1

Characteristics of 13 Psychometric Marker Scales Representing Diverse Favorability-Endorsement Combinations as They Occur in the MMPI

Scale Characteristic	Keying	Favorability Rating	Endorsement Percentage	No. of Items	Label
Extreme Deviance	True	1.0-2.9	00-19	68	ED_t
	False	7.0-8.9	80-100	30	ED_f
Moderate Deviance	True	2.5-3.9	20-39	56	MD_t
	False	6.0-7.4	60-79	40	MD_f
Obvious Favorability	True	6.0-8.9	00-49	11	OF_t
	False	1.0-3.9	50-100	9	OF_f
Slight Favorability	True	6.0-8.9	50-59	9	SF_t
	False	1.0-3.9	40-49	13	SF_f
Communality	True	4.5-5.4	60-100	29	C_t
	False	4.5-5.4	00-39	38	C_f
Slight Communality	True	5.5-5.9	60-100	33	SC_t
	False	4.0-4.4	00-39	38	SC_f
"Acquiescence"	True	4.0-5.9	40-59	51	A

From the foregoing procedures 13 mutually exclusive "scales" were developed which possess the characteristics presented in Table 1. A subject's score on such a "scale" is simply the number of

times his answers to the items in a given endorsement-desirability category agree with the predetermined scoring key for that category. The scales were scored for a sample of 250 Stanford University undergraduate men who had taken the complete MMPI under standard instructions.²

The intercorrelation matrix of the 13 scales was submitted to a principal axis factor analysis, using unities as communality estimates and retaining only factors with eigenvalues greater than one (Harman, 1960). The resulting factor matrix was analytically rotated to a varimax criterion of simple structure (Kaiser, 1958).³

Results

Table 2 presents the rotated factor matrix of 13 Psychometric Marker scales based on 250 men. The three rotated factors account for approximately 69 percent of the total possible variance among the 13 original scales. The first factor is a large one and accounts for some 52 percent of the common variance among the scales. The remaining two factors account for 28 and 20 percent of the remaining common variance.

TABLE 2
Rotated Factor Matrix of 13 Psychometric Marker Scales
(*N* = 250 college men)

Scale Characteristic	Keying	Label	I	II	III	<i>h</i> ²
Extreme	True	ED ₁	60	61	-11	75
Deviance	False	ED ₂	18	73	24	63
Moderate	True	MD ₁	83	39	15	86
Deviance	False	MD ₂	21	56	58	71
Obvious	True	OF ₁	-04	-02	-79	63
Favorability	False	OF ₂	-79	-03	-21	67
Slight	True	SF ₁	-05	-29	-60	45
Favorability	False	SF ₂	-82	-13	-17	72
Communality	True	C ₁	72	-41	12	70
	False	C ₂	-60	-40	48	76
Slight	True	SC ₁	14	-75	-17	61
Communality	False	SC ₂	-83	-36	00	83
"Acquiescence"	True	A	79	-14	-24	71
%Common Variance			52%	28%	20%	

² The cooperation of Ernest R. Hilgard in making available some of the MMPI protocols is gratefully acknowledged.

³ A scale intercorrelation matrix and an unrotated factor matrix have been deposited with the American Documentation Institute. Order Document No. 8317 remitting \$1.25 for 35 mm. microfilm or \$1.25 for 6 by 8 photocopies.

Factor I

The first factor is a general factor in that at least one member of each of the classifications loads it substantially. *Deviance* subcategories are all positively loaded and have significant loadings on the True subscales only. Conversely, *Favorability* categories are all negatively loaded and have significant loadings on the False subscales only. True and False subcategories of *Communality* load in opposite directions and, in the case of the Communality scales C_t and C_f , substantial loadings occur with curiously reversed signs. Were we to compute factor scores for individuals, subjects achieving a high score on this factor would do so by answering "true" to items of diverse psychometric properties. However, this "preference" for "true" answering must be recognized as being relatively independent of tendencies to answer false to what might be considered as psychometric reversals of the items.

For Deviance and Favorability, the situation seems clear enough. A hypothetical respondent would score high on this factor by answering "true" to deviant items of low favorability-endorsement and "true" to items of low favorability with varying but generally moderate endorsement. Tendencies to answer "false" to the psychometrically-opposite categories of high favorability-endorsement and high favorability with moderately low endorsement are better explained in terms of another orthogonal factor (i.e., Factor III). The statistical independence of what have been termed Deviant True and Deviant False has been previously noted (Barnes, 1956; Wiggins, 1962c; Block, 1962). The possible independence of Unfavorable True and Unfavorable False has previously been only suspected (Kogan and Boe, 1961; Schultz, Kogan, and Chapman, 1962).

Where desirability values are in the neutral range, our hypothetical respondent appears to answer "true" regardless of the level of item endorsement (C_t and C_f). Edwards (1961; Edwards and Dires, 1962) has suggested that neutrality of desirability values provides an occasion for ambiguity in which acquiescent tendencies may emerge. A critical test of the relative contributions of endorsement and favorability within the MMPI is made difficult by the lack of items of both neutral favorability and extreme endorsement.

The substantial and positive loading of the "Acquiescence" scale on this factor suggests that other things being equal (and indifferent), high scorers on this factor will answer "true." Since the "acquiescence" implied by this scale has been held to be independent of "social desirability" (Edwards, 1961; Jackson and Messick, 1961) and, to some extent, of Deviant True (Wiggins, 1962c), its appearance on this factor requires further comment. In the original *unrotated* factor matrix, A had approximately equal and substantial loadings on both the first and second factors (.63 and $-.56$, respectively). Such a position is compatible with the interpretation of acquiescence as a "fusion factor" of Deviant True and Deviant False (Wiggins, 1962c). However, the extent to which acquiescence loads one or the other of the two main factors of the MMPI is often a matter of rotational preference (Messick and Jackson, 1961a; Edwards and Walker, 1961). The present rotation was performed analytically without concern for the optimal positioning of acquiescence.

The mutual and complex contributions of Deviant True, Unfavorability True, and "Acquiescence" to a single, general, and important (35 percent of total variance) factor of MMPI scale intercorrelations give rise to a strong feeling of *deja vue*. Clearly this is related to the highly confounded (and, as Block (1962) quips, "confounding") first factor of the MMPI which has recently flown under the neutral colors of "Alpha." Confirmation of this suspicion is achieved later in this paper.

Factor II

The second factor is dominated by Deviance and Communality Categories in the absence of significant contributions from any of the categories of Favorability. The signs of the loadings indicate a factor characterized by the tendency to answer items in the direction of lesser groups consensus (non-communality) *whether this direction be keyed true or false*. This is accomplished by choosing unpopular response options, both when these options are undesirable and when they are indifferent with respect to desirability. The insignificant loadings of the Favorability categories suggest the relative *independence* of communality-responding from favorability-responding. (The signs of the Favorability categories are nonetheless consistently negative, as would be expected.) It seems

likely that this factor is related to the usually obtained second factor of the MMPI which Block (1962) has designated as "Beta."

Factor III

The third factor has substantial loadings on scales in which item desirability exceeds item endorsement (OF_t , SF_t , C_t). Loadings are negative when keying is in the desirable direction and positive when keying is in the popular direction. A hypothetical respondent high on this factor would tend to respond "false" when desirability exceeds endorsement; one who is low tends to respond "true." Only the substantial loading of MD_t cannot be subsumed under this description. Although little has been written about the recently isolated third factor of the MMPI (Edwards, Diers, and Walker, 1962; Wiggins, 1964b) the present factor is tentatively labelled "Gamma" in a hope for consistency.

Factor Identification

A brief summary of procedures seems indicated at this point. A scatter plot of the item characteristics of rated favorability and empirical endorsement was used as a basis for constructing scales which were representative of diverse favorability-endorsement combinations that occur in the MMPI. This was done mechanically without reference to the "content" of the items or their membership in clinical or any other existing scales of the MMPI. Intercorrelations among these scales were factor analyzed and rotated by mechanical analytic procedures. Lacking any substantive referents, the factor matrix was "interpreted" in terms of response patterns to the item characteristics of endorsement and favorability. Factor Alpha is a general and rather complicated factor for which we invoked hypothetical response tendencies of Deviant True, Unfavorability True, and Acquiescence. Factor Beta was interpreted as involving responses to unpopular item options (non-communality) regardless of the direction of item keying. Factor Gamma suggested the operation of an Unfavorability False pattern of response particularly to items for which desirability values exceeded endorsement percentages.

The above factor interpretations were, at various points, reminiscent of the three factors that have been identified in studies involving more familiar clinical and stylistic scales of the MMPI. To

subject this intuition to a direct test, the best marker scale was selected from each of the three psychometric factors just described and included in a correlation matrix of more familiar MMPI scales. The "best" marker of each factor was assumed to be that scale which had the highest loading on the factor to be marked and the lowest loadings on the other two factors. From Table 2 it can be seen that, in the above sense, SF₁ marks Factor Alpha, SC₄ marks Factor Beta, and OF₁ marks Factor Gamma.

The scales selected for inclusion in the marker study were the familiar MMPI clinical and validity scales, Welsh's (1956) Factor Scales *A* and *B*, and four stylistic response scales that have been demonstrated to be relatively unique markers of the three factors of the MMPI. The stylistic scales are: Edwards (1957) *SD* (Factor Alpha), Wiggins (1962b) *Rb* (Factor Beta), Cofer et al.'s (1949) *Mp*, and Wiggins (1959) *Sd* (Factor Gamma scales). Evidence supporting the inclusion of these scales as reference points for the factorial structure of the MMPI may be found in Edwards, Diers, and Walker (1962) and Wiggins (1964).

Correlations were obtained among the 22 clinical, validity, stylistic, and psychometric marker scales in the group of 250 Stanford University male students. The resultant intercorrelation matrix was submitted to a principal axis factor analysis with unities in the diagonal and factors with eigenvalues greater than one were then rotated to a Varimax criterion.⁴ The four factors obtained by these procedures accounted for approximately 69 percent of the total possible variance among the scales. Table 3 presents the rotated factor matrix of 22 scales. Scales are ordered so as to highlight the factor structure and loadings with absolute values less than .33 are omitted in the interest of clarity.

Inspection of Table 3 suggests that the first two factors of the present analysis are similar to those that have been consistently isolated in earlier studies of this instrument (e.g., Wheeler, Little, and Lehner, 1951; Welsh, 1956) and the first three factors are typical of more recent studies that have included stylistic scales (Edwards et al., 1962; Wiggins, 1964). The fourth factor, which is not relevant to the present analysis, appears to combine Eichman's (1961, 1962)

⁴ A scale intercorrelation matrix and complete rotated and unrotated factor matrices have been deposited with A.D.I. (see footnote 3).

TABLE 3

Rotated Factor Matrix of MMPI Scales and Three Psychometric Markers
(*N* = 250 college males)

	I	II	III	IV	<i>h</i> ²
SD	-92				87
A	90				89
Pt	88				89
K	-87				84
Sc	83			-36	87
Si	81	33			83
D	58	39			63
Hs	55			-46	56
F	54			-47	54
Mf	41			-40	40
Pd	39			-61	60
Pa				-64	45
Hy				-82	74
R		81			71
Rb		-73			58
Ma	34	-60			59
L		55	55		64
Sd			84		81
Cof	-33		79		74
(I) SF ₁	-75	36			73
(II) SC ₁	36	-66			61
(III) OF ₁			75		61
% Common Variance	45%	21%	17%	17%	

third (somatization) and fourth (unconventionality) factors which have appeared from time to time in other analyses (e.g., Fisher, 1957; Kassebaum, Couch, and Slater, 1959).

The loadings of the three psychometric marker scales lend support to the hypothesis that the three factors isolated in the original analysis of psychometric scales are congruent with the three factors which have emerged in studies employing more familiar MMPI clinical and stylistic scales (Edwards et al., 1962; Wiggins, 1964). Psychometric Scale SF₁ clearly loads the first factor, SC₁ loads the second, and OF₁ loads the third factor. The size of the loading of each marker scale on its appropriate factor is substantial and cross factor loadings, though present between Factors I and II, are not large. Although an even more dramatic fit might be possible under a different rotational procedure, analytic rotation in terms of the more heavily represented MMPI clinical scale variance was more germane to the hypothesis under test.

Discussion

It has been demonstrated that the three principal dimensions of the MMPI may be reproduced by scales which were constructed solely on the basis of the item characteristics of communality and favorability. Although the importance of these two item characteristics as sources of method variance in the MMPI is thereby highlighted, it should be strongly emphasized that such a demonstration does not minimize the possible contribution of other sources of variance of a psychometric or substantive nature. Psychometric and substantive characteristics of the 566 statements that constitute the MMPI are so greatly confounded that scale construction in terms of any one major characteristic will almost inevitably involve other characteristics as well (Wiggins, 1962c; Block, 1962).

Edwards and Heathers (1962) have argued against a substantive interpretation of the first factor of the MMPI because of the demonstrated contribution of favorability values to this dimension. Jackson and Messick (1961) have minimized the substantive potential of the second factor of the MMPI on the basis of their demonstration that item keying is a potent source of variance in this factor. Block (1962) has attempted to refute the arguments of these investigators by demonstrating the reproducibility of the familiar MMPI factors even after the psychometric influence of favorability has been removed from the first factor and opportunities for acquiescence removed from the second. The results of the present study are not meant to reopen the issues Block attempted to close. It may not be possible to choose between substantive and stylistic interpretations of the principal dimensions of the MMPI when the item pool is so highly confounded in this respect. Block's (1962) plea for parsimony is appealing in light of the complexities of inference demanded by the present method of analysis. Nevertheless, it is unlikely that substantive interpretations of the MMPI will ever be uncritically accepted until the contribution of psychometric characteristics is better understood. The present study suggests that the joint characteristics of communality and favorability must be reckoned with as potent, although not clearly understood, sources of method variance which contribute to the structure of the MMPI.

Summary

A scatter plot of the item characteristics of rated favorability and endorsement percentage was used as a basis for constructing scales which were representative of diverse favorability-endorsement combinations that occur in the MMPI. Intercorrelations among these scales were factor analyzed and rotated by mechanical analytic procedures. The resultant three factors were reminiscent of the three factors that have been identified in studies involving more familiar MMPI clinical and stylistic scales.

Inclusion of three scales, which uniquely loaded the three factors, in a larger factor analysis involving MMPI clinical and stylistic scales tended to confirm the similarity of the original factors to the more familiar ones. Although the importance of the item characteristics of communality and favorability was highlighted by this study, the results were not interpreted as minimizing the possible contribution of other sources of variance. The highly confounded nature of the MMPI item pool is such that substantive interpretations of the principal dimensions of the test must await further clarifications of the contribution of method variance to the structure of the instrument.

REFERENCES

- Barnes, E. H. "Response Bias in the MMPI." *Journal of Consulting Psychology*, XX (1956), 371-374.
- Block, J. "Unconfounding Meaning, Acquiescence and Social Desirability in the MMPI." Unpublished manuscript, Institute of Human Development, University of California, Berkeley, 1962.
- Cofer, C. N., Chance, June, and Judson, A. J. "A Study of Malingering on the MMPI." *Journal of Psychology*, XXVII (1949), 491-499.
- Edwards, A. L. "The Relationship between the Judged Desirability of a Trait and the Probability That the Trait Will Be Endorsed." *Journal of Applied Psychology*, XXXVII (1953), 90-93.
- Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden Press, 1957.
- Edwards, A. L. "Social Desirability or Acquiescence in the MMPI? A Case Study with the SD Scale." *Journal of Abnormal and Social Psychology*, LXIII (1961), 351-359.
- Edwards, A. L. and Diers, Carol J. "Social Desirability and the Factorial Interpretation of the MMPI." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 501-509.
- Edwards, A. L., Diers, Carol J., and Walker, J. N. "Response Sets

- and Factor Loadings on Sixty-One Personality Scales." *Journal of Applied Psychology*, XLVI (1962), 220-225.
- Edwards, A. L. and Heathers, Louise B. "The First Factor of the MMPI: Social Desirability or Ego Strength?" *Journal of Consulting Psychology*, XXVI (1962), 99-100.
- Edwards, A. L. and Walker, J. N. "Social Desirability and Agreement Response Set." *Journal of Abnormal and Social Psychology*, LXII (1961), 180-183.
- Eichman, W. J. "Replicated Factors on the MMPI with Female NP Patients." *Journal of Consulting Psychology*, XXV (1961), 55-60.
- Eichman, W. J. "Factored Scales for the MMPI." *Journal of Clinical Psychology*, (1962), Monograph Supplement 15.
- Fisher, J. "An Empirical Study of the Relation of Physical Disease to Body-Object Cathexis." Unpublished manuscript, VA Hospital, San Francisco, California, 1957.
- Goldberg, L. R. and Rorer, L. G. "Test-Retest Item Statistics for Original and Reversed MMPI Items." *Oregon Research Institute Research Monograph*, III, No. 1 (1963), Eugene, Oregon.
- Harman, H. *Modern Factor Analysis*. Chicago: University of Chicago Press, 1960.
- Jackson, D. N. and Messick, S. "Acquiescence and Desirability as Response Determinants on the MMPI." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 771-790.
- Kaiser, H. F. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika*, XXIII (1958), 187-200.
- Kassebaum, G., Couch, A. S., and Slater, P. E. "The Factorial Dimensions of the MMPI." *Journal of Consulting Psychology*, XXIII (1959), 226-236.
- Kogan, W. S. and Boe, E. E. "The Social Desirability Variable: Differential Response to High and Low Value Items." Paper read at the Western Psychological Association meeting, Seattle, Washington, June, 1961.
- Messick, S. and Jackson, D. N. "Acquiescence and the Factorial Interpretation of the MMPI." *Psychological Bulletin*, LVIII (1961), 299-304. (a)
- Messick, S. and Jackson, D. N. "Desirability Scale Values and Dispersion for MMPI Items." *Psychological Reports*, VIII (1961), 409-414. (b)
- Schultz, C. B., Kogan, W. S., and Chapman, H. "Favorability, Unfavorability, and Content Considerations in SD Scales." *Psychological Reports*, X (1962), 619-622.
- Welsh, G. S. "Factor Dimensions A and R." In G. S. Welsh and W. G. Dahlstrom (Editors), *Basic Readings on the MMPI in Psychology and Medicine*. Minneapolis: University of Minnesota Press, 1956.
- Wheeler, W. M., Little, K. B., and Lehner, G. F. J. "The Internal Structure of the MMPI." *Journal of Consulting Psychology*, XV (1951), 134-141.
- Wiggins, J. S. "Interrelationships among MMPI Measures of Dis-

simulation under Standard and Social Desirability Instructions." *Journal of Consulting Psychology*, XXIII (1959), 419-427.

Wiggins, J. S. "Components of Variance in Objective Personality Tests." Paper presented at the annual meetings of the Western Psychological Association, San Francisco, California, April 20, 1962. (a)

Wiggins, J. S. "Definitions of Social Desirability and Acquiescence in Personality Inventories." In S. Messick and J. Ross (Editors), *Measurement in Personality and Cognition*. New York: John Wiley & Sons, 1962. (b)

Wiggins, J. S. "Strategic, Method and Stylistic Variance in the MMPI." *Psychological Bulletin*, LIX (1962), 224-242. (c)

Wiggins, J. S. "An MMPI Item Characteristic Deck." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIV (1964), 137-141. (a)

Wiggins, J. S. "Convergences among Stylistic Response Measures from Objective Personality Tests." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIV (1964), 551-562. (b)

DIFFERENTIAL CONTENT VALIDITY: THE CALIFORNIA SPELLING TEST, AN ILLUSTRATIVE EXAMPLE

KENNETH D. HOPKINS

University of Colorado

AND

CAROLYN J. WILKERSON

Arizona State University

"In judging the content validity of an achievement test, the first question is whether the test covers a representative sample of the curricular content" (Anastasi, 1961, p. 437). "Thus the content validity of the geography test would have to be studied by checking the items against the course of study the students have followed" (Cronbach, 1960, p. 104). As explicitly indicated by the above quotations, the items on an achievement test must "mirror" the curricular objectives of a given course of study in order to possess content validity. School districts have been urged to select the standardized achievement measures which are most congruent with their particular objectives. Such a procedure, however, presents a serious problem of interpretation. If the test with the greatest content-curriculum parallelism is selected, how can it be evaluated by norms based on schools with random emphasis and degrees of parallelism, i.e., by norms from a *representative* sample of schools? If two districts are equal in every respect, including instructional effectiveness, yet have differing curricular objectives, can they legitimately be compared on a standardized test which reflects objectives almost certainly to be more relevant for one of the districts?

The problem of norm interpretation resulting from differential content validity does not seem to have received adequate attention.

None of the current books on measurement consulted, Adams, Ahmann and Glock, Anastasi, Cronbach, Davis, Downie, Durost and Prescott, Freeman, Furst, Gerberich et al., Gronlund, Guilford, Gulliksen, Helmstadter, Lindquist, Lindvall, Noll, Nunnally, Stanley, Remmers et al., Thorndike and Hagen, made any mention of the problem.

If a standardized achievement test includes only those curricular areas that are "common" to all districts, is a "fair" picture given of the more flexible districts which have introduced many additional supplementary concepts into their instructional program? The more stereotyped district might, according to the district means, appear to have a more effective instructional program since more time is spent on test-relevant materials.

Some studies have attempted to evaluate equivalence of norms by giving comparable groups various achievement tests and examining resulting mean grade equivalents. Stake (1961, p. 61), for example, concluded that the California Achievement Test (CAT), "... yields grade placements that are significantly higher than those from four other widely used test batteries." Although his conclusion may be valid, it was not proven by his study since no consideration was given to the possibility of differential content validity, i.e., that the CAT may just have had more curricular relevance for the particular locality in which the study was conducted. A study in a different geographical area at different grade levels did not reveal such a marked consistent trend (Garlock, 1959). Even a cursory examination of the various achievement tests having the same label will reveal that they are not parallel in emphasis or content. Unfortunately Stake (1961) did not indicate whether his study dealt with a particular subject area, such as reading, or with the combined achievement picture. Garlock (1959) found wide differences depending on the specific subject in question.

Although there have been a few studies (Howell and Weiner, 1961; Stake, 1961; Taylor, 1961) that relate to the problem of differential content validity, the investigators seem to interpret the differences as norm differentials, overlooking the possibility of differences in curricular relevance. Almost no attention has been given to an evaluation of the possibility of differential content validity between alternate forms of a given achievement series. Stake (1961) made no mention of the forms of the tests involved in his study, yet

made a conclusion generalized to all forms, apparently assuming that all forms had equal curricular relevance.

Method. The present study purposed to determine if significant differential content validity existed on the California Spelling Test, one of six content tests on the CAT battery. This particular test was selected since four alternate forms are available which would facilitate a more sensitive assessment of differential content validity. The subject of spelling was chosen since it lends itself to an unambiguous, objective analysis.

The four forms, W, X, Y, Z, of the test were evaluated for equivalence by relating them to the course of study guide used by the State of California (Madden and Carlson, 1956). The format of the tests calls for the examinee to identify the misspelled word (if any) from the four given for each item. If the misspelled word is correctly recognized, he is credited with the item. Grade equivalents were determined assuming an examinee correctly recognized only those misspelled words that have been formally introduced to him in his "speller." These "curriculum oriented" grade equivalents for the four forms of the elementary battery in respective ordinal sequences were:

4th grade: 3.9, 3.1, 2.8, 2.6 for X, Y, Z, W; a range of 1.3;

5th grade: 4.9, 4.7, 4.3, 3.7 for X, Z, Y, W; a range of 1.2;

6th grade: 7.0, 6.6, 5.5, 4.9 for X, Y, Z, W; a range of 2.1.

It is interesting to note that the lowest values occurred at each grade on the most widely used form, W; in fact, only form W is approved for the mandatory statewide testing in California at grade five.

Using the junior high level of the CAT spelling test, the following grade equivalents resulted:

7th grade: 6.4, 6.3, 6.1, 5.8 for X, Y, Z, W; a range of 1.1;

8th grade: 8.3, 7.5, 7.5, 6.9 for X, Y, W, Z; a range of 1.4;

9th grade: 10.0, 10.0, 9.3, 8.4 for W, Z, X, Y; a range of 1.6.

Although these findings are suggestive evidence for differential content validity among forms, no definitive conclusions can be drawn. In addition to determining whether a test corresponds with specific spelling curriculum, one must also see if the items have differential difficulty resulting from the course of study as well. Theoretically it would be possible for the words appearing in a course of

study to have been selected not only to be learned per se, but also to illustrate certain principles and rules of spelling that enable the student to generalize to words of a similar structure. One cannot conclude that just because one form of a standardized test contains more words that have been taught to a given group of students, that they will necessarily score higher on that form. Perhaps the words of the alternate forms are actually perfectly parallel, i.e., from knowledge about a given word—configuration, phonetic combination, etc.—other words can be spelled through generalization. Certainly one would expect a fifth grade child who has been taught to spell “battle,” also be able to spell “rattle” which has not been taught.

There is another factor which makes the previously presented tabulational analyses equivocal. It goes without saying that words differ in difficulty, consequently, one could not conclude that grade equivalent scores on alternate forms would differ just because the forms happen to differ in the number of words they contain which have been formally taught. The form with fewer taught words might contain words at a lower difficulty level which would compensate for the instructional variance.

To obtain empirical evidence regarding any differential content validity between forms, the four alternate tests at the elementary level were examined, with all items being categorized into one of the following four groups:

1. words appearing correctly spelled on the test that have appeared in the curriculum;
2. words appearing correctly spelled on the test that have not appeared in the curriculum;
3. words appearing misspelled on the test that have appeared in the curriculum; and
4. words appearing misspelled on the test that have not appeared in the curriculum.

The comparisons that are relevant for the study are: 1 with 2, 1 with 4, and 1 and 3 with 2 and 4. A special test was devised to allow these comparisons. The absolute difficulty of the words was controlled using the publisher's criterion of item placement. The test employed a difficulty gradient, therefore, each “taught” and “not taught” word was matched by item number. For example, if “matter” has appeared in the curriculum and was found at item number

in a misspelled form, then, from the other forms, all misspelled words appearing in item number 20 that have not appeared in the curriculum were placed in a box and one was drawn randomly to appear on the experimental test. The word "master" would be in category "3" and its paired "untaught" word in category "4."

All fifth-grade students ($N = 139$) in three elementary schools in Los Angeles County (mean CTMM IQ = 104) were administered the 56 word experimental spelling test; 14 words from each of the four categories, matched by item number. Written reproduction of the dictated words was required of the examinees; each word was pronounced, used in a sentence, and pronounced again. The dictation format was used since it is the most criterion-oriented approach and also avoids the possible contamination from response sets and chance.

Results. The "taught" words were spelled correctly significantly more often than the "untaught" words of equal absolute difficulty as revealed in Table 1. Of the 28 words that had appeared at some time during the first five years of formal schooling, 63.6 percent

TABLE 1

Means, Standard Deviations, Reliability Coefficients, Intercorrelations, and t-ratios for the Experimental Spelling Tests Extracted from the Alternate Forms of the CAT Spelling Tests

Word Format Used on CAT Test						
	Misspelled		Correctly Spelled		Totals	
	Taught	Untaught	Taught	Untaught	Taught	Untaught
Number of Words	14	14	14	14	28	28
\bar{X}	8.04	5.42	9.78	6.07	17.82	11.50
s	4.15	3.38	3.58	3.83	7.50	7.00
r (#21)	.87	.77	.83	.83	.92	.90
	4.78**		5.25**		18.27**	
Intercorrelations						
Misspelled						
Taught			.83	.88	.83	.98
Untaught				.75	.88	.81
Correctly Spelled-Taught					.77	.97
Correctly Spelled-Untaught						.83
Totals-Taught						.97
						.85

** $p < .0001$.

*Appreciation is expressed to the Western Data Processing Center for making computer time on the IBM 7090 available to the writer under its cooperative plan of institutional participation in non-profit research activities.

were correctly spelled on the dictation test; of the 28 that had not been taught, only 41.1 percent were correctly spelled. The difference was highly significant (.0001 level). As would be expected, the results were the same regardless of the form (misspelled or correctly spelled) in which the words appeared on the standardized tests.

The results indicate that the formal teaching of a word did decrease its relative difficulty; it is also apparent, then, that a California school district would attain a higher grade placement average if they used the form with the largest number of "taught" words appearing in it, i.e., form X at grade five. Differential content validity did exist on these tests, resulting from the variance in curricular parallelism found between forms. The differential converted into grade equivalents was estimated to be 2 to 3 school months. This estimate was derived by computing the difficulty differential between the "taught" and "untaught" words and applying this value in relation to the actual numbers of words in these categories on test forms W and X at the elementary level. Although the importance in the interpretation of individual scores is apparent, much greater significance relates to the difficulty introduced in district or school evaluations in which a differential of two or three months becomes a major interpretive problem.

Summary. The study focused upon an unattended problem in standardized achievement testing, viz., differential content validity and its concomitant interpretive problems. It would appear that greater attention should be given to curricular-test relevance in the construction, comparing, selecting, and interpreting of tests—even within the same achievement series. The present study found differential content validity even among the alternate forms of a given test. The phenomenon found with the CAT spelling tests is not unique (Howell and Weiner, 1961), and warrants further investigation with other tests and curricula.

REFERENCES

- Anastasi, Anne. *Psychological Testing* (Second Edition). New York: Macmillan, 1961.
- Cronbach, Lee J. *Essentials of Psychological Testing* (Second Edition). New York: Harper and Brothers, 1960.
- Garlock, Gerald C. "A Study Comparing the California Achievement Test, the SRA Achievement Series, the Sequential Tests of Edu-

- cational Progress, and the Stanford Achievement Series." Los Angeles County Superintendent of Schools, Division of Research and Guidance, 1959.
- Howell, John J. and Weiner, Max. "Note on the Equivalence of Alternate Forms of an Achievement Test." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 309-313.
- Madden, Richard and Carlson, Thorsten. *Teaching for Success in Spelling*. (Workbooks for grades 2-8). New York: World Book Company, 1956.
- Stake, Robert E. "'Overestimation' of Achievement with the California Achievement Test." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 59-62.
- Taylor, Edward A. and Crandall, James H. "A Study of the 'Norm Equivalence' on Certain Tests Approved for the California State Testing Program." *California Journal of Educational Research*, XIII (1961), 186-192.

THE EPPS PATTERN AND THE "NURSING PERSONALITY"¹

DANIEL V. CAPUTO² AND CONSTANCE HANF³

Washington University School of Medicine

It is a widely held tenet that personality needs or attitudes are strongly involved both in vocational (or professional) choice (Kuder, 1951) and in vocational continuance (Kibrick, 1958). Strong (1951, p. 3) stated that, "Women engaged in a particular occupation have a characteristic pattern of likes and dislikes, which differentiate them from women following other professions."

Further, the hypothesis that personality needs heavily influence vocational selection necessarily presupposes that there is a set of attitudes or personality needs shared in common by those *already functioning in the given vocation* and that individuals seeking to enter this group also share this common set of needs.

The present paper assesses this hypothesis in the field of nursing. According to this view one would expect that there is a representative pattern of personality needs among nurses which is shared in common by students already in nursing education and by those who have chosen nursing as a career. It would be further assumed that this pattern is meaningful enough to discriminate those in the nursing area from non-nursing groups despite the diversity of tasks

¹ This paper is a partial report of a research project "Role Differential and Nursing Ideology," John A. Stern and Albert F. Wessen, Co-Principal Investigators; the research has been supported by National Institute of Health Grant NU-00050. A portion of the computation was done with support provided by the Washington University Computer Center under National Science Foundation Grant G-22296.

² Now at Queens College of the City University of New York.

³ Now at the University of Oregon Medical School. The authors wish to express their gratitude to the co-principal investigators, to Jon M. Plapp, and to Dr. George Psathas for their invaluable suggestions and aid.

and particular situations subsumed under the rubric "nursing."

If the EPPS can be assumed to be a reasonable measure of such patterns of personality needs, then 1) intercorrelations of overall EPPS patterns among diverse nursing and nursing-student aggregations ought to show a significant positive relationship, and 2) intercorrelations of EPPS patterns among nursing groups ought to be significantly higher than those between nursing and non-nursing groups. Confirmation of these hypotheses would add support to the view that personality factors are highly related to occupational choice. Within the purview of the present study, there are four possible conceptions relating personality need pattern and occupational or professional choice. These are graphically summarized in Table 1. Number 1 in Table 1 represents the conception just described.

TABLE 1

Conceptions Relating Personality-Pattern and Occupational Choice for this Study

<i>Entrants into Profession</i>		<i>Professionals</i>
1. Common Need Pattern	—————→	Common Need Pattern
2. Random Need Pattern	—————→	Common Need Pattern
3. Random Need Pattern	—————→	Random Need Pattern
4. Common Need Pattern	—————→	Random Need Pattern

A second mode of viewing occupational choice might stipulate that factors unrelated to personal needs or attitudes are primary determinants. Such a view would hold that extra-personality influences such as finances, opportunity, and others are preeminent. The implication of this view is that those entering into training for a particular profession or occupation present a random pattern of attitudes or personality needs which might then be molded into a pattern characteristic of the professional group.

A third view might hold that groups of individuals entering a profession present a random attitude pattern which remains random for the group as a whole. This conception (Number 3 in Table 1) would discard the notion that members of an occupational group share or develop a common pattern of personality needs. It suggests that an entity as broad as an "occupation" or "profession" has as much variability within it as between it and other occupations or professions. A "profession" then, would be seen as too molar a concept to have any great relevance to the personality-need patterns of unselected groups subsumed under it.

A fourth conception might stipulate that groups of individuals *seeking entrance* into a particular profession would present a similar need pattern to one another based perhaps on some global, popular idea of the profession. However, this need pattern would be differentially modified by diversity of training and experience within the profession, leading to random, group patterns.

The EPPS has often been employed to assess the personality-need patterns of nurses and nursing students.

Williamson, Edmonston, and Stern (in press) found that a group of nurses from various services of a single hospital-complex differed significantly (*t* tests) from the Edwards (1959) norms for college women on 10 of the 15 EPPS needs, while differing from each of three groups of Veterans Administration nurses obtained by Navran and Stauffacher (1957) on only three of the 15 needs. Their findings pointed up the possibility that there is a hierarchy of EPPS needs peculiar to the nursing profession as a whole and to nursing specialties in particular.

Navran and Stauffacher discovered (1957) that a subgroup of VA psychiatric nurses differed significantly from college women on seven of the EPPS needs (*t* tests). Six of these seven needs, order, deference, endurance, affiliation, autonomy, and exhibitionism, were the same as those which differentiated the group of nurses from college women in the study by Williamson et al.

Gynther and Gertz (1962) obtained the EPPS profiles of advanced nursing students from most of the South Carolina nursing schools. They found that these students were significantly different from *either* college women or high school girls on six of the EPPS needs, four of these, order, exhibitionism, autonomy, and endurance being the same as those found by Navran and Stauffacher, and Williamson, et al.

Zuckerman (1958) in his study employing sophomore nursing students found 5 EPPS needs which differentiated them from Edwards' college-women controls. Only one of these, autonomy, was consistent with the Williamson, et al., Navran and Stauffacher, and Gynther and Gertz findings.

Redden and Scales (1961) noted 12 significant differences between their group of Negro nursing students and college women. They pointed out in addition, that nursing education was not influential in effecting important changes in the individual's need

system since freshman and senior students differed on only 2 EPPS variables.

While these studies indicate the possibility of a unique "nursing personality" as measured by the EPPS, others cast doubt on this possibility.

Reece (1961) in a study concerned primarily with the personality-need characteristics of successful and unsuccessful nursing students noted, "In terms of the rank order of these (needs), the *Ss* of this study appear to be more similar to Edwards' norm group than to Navran and Stauffacher's *Ss*" (p. 174). Similarly, Healey and Borg (1951), employing the Guilford-Martin battery of personality tests, were unable to find a characteristic personality pattern for beginning nurses.

Thus, the question of whether there is a personality-need profile characteristic of nurses, independent of locale or subspecialty, which is shared by those entering the profession (and therefore possibly involved in vocational choice) remains an open one.

Method

The methodology of the present study involved obtaining a measure of the personality-need patterns of a series of nursing groups for comparison with one another and with those of comparable non-nursing groups.

Instrument

The EPPS was the instrument employed in this study because

TABLE 2
Subjects Employed

Comparison Group	N	Reference
RN's General Hospital I (GH I), St. Louis	50	Present study
RN's General Hospital II (GH II), St. Louis	35	Williamson et al.
Freshman nursing students, General Hospital I (GH I) St. Louis	79	Present study
Freshman nursing students, General Hospital III (GH III) Detroit	87	Reece
Senior nursing students, General Hospital I (GH I) St. Louis	62	Present study
Senior nursing students, General Hospital IV (GH IV) South Carolina	222	Gynther and Gertz*
Adult women	4,932	Edwards
College women	749	Edwards
High school girls	834	Klett

*"Advanced students" assumed to be seniors.

TABLE 3
Background and Professional Characteristics of Two RN Groups Sampled

	Mean Age (Rounded)	N & (%) Married	N & (%) with Children	N & (%) Negro	N	Nursing Service			
						Rehab.	Obstet.	Surg. Med.	Med.
GH I									
RN's	27	18 (36)	13 (26)	7 (14)	50	1	17	3	14
GH II									
RN's	28	13 (37)	5 (14)	8 (23)	35	0	10	0	11
Chi									
Squared*	0	0	1.69	1.11					

*None was significant.

1) it is a relatively efficient way of sampling large numbers of Ss, and 2) it has been extensively employed to delineate group "personality needs" in nursing as well as non-nursing groups so that comparison groups were available. In all cases, the EPPS was administered in the standard manner.

In order to include nursing groups from different locales and at different levels of training both the EPPS results from some of the studies cited and EPPS results obtained specifically for the present study, were employed in testing the hypotheses mentioned.

The subject groups employed in this study are described in Table 2.

Table 3 compares the background and professional characteristics of the two registered nurse (RN) groups.

Procedure

The procedure entailed comparing the EPPS patterns of nursing groups among themselves and with appropriate controls (see Table 2). EPPS raw scores were ranked within each group and Spearman rhos (rank correlations) were computed. (See Table 4 for the rankings of the EPPS by the subject groups.) In this way, groups for which only mean scores for EPPS needs were available, could be employed in the comparisons. The ranking method allows one to focus on molar, pattern comparisons as well as on specific need

TABLE 4
EPPS Need Rankings (Based on Raw Scores)

	RN's GH I	RN's GH II	Seniors GH I	Seniors GH IV	Fresh- men GH I	Fresh- men GH III	College Women	Adult Women	HS Girls
Ach	11	9	9	14	12	11	9	9	14
Def	10	7	15	9	9	9	12	8	12
Ord	13	8	14	11	14	14	15	6	15
Exh	9	11	7	8	6	6	7	12	6
Aut	15	15	11	13	13	12	11	11	11
Aff	4	6	6	4	5	4	1	2	2
Int	1	1	5	1	1	1	2	7	5
Suc	12	14	8	10	8	13	11	10	8
Dom	8	10	10	12	11	10	8	13	9
Afa	5	5	4	3	3	5	5	3	3
Nur	2	4	3	2	2	2	4	1	4
Chg	3	3	2	5	4	3	3	5	1
End	7	2	12	7	10	7	10	4	10
Het	6	12	1	6	7	8	6	15	7
Agg	14	13	13	15	15	15	14	14	13

rankings. The fact that it discounts the number of subjects involved (since N is always 15, the number of EPPS needs) may be a liability, however, since it forces the assumption that one set of rankings is as reliable as another. In addition, material is dealt with on an ordinal basis, some degree of discrimination being sacrificed.

Borislow (1958) noted that the EPPS consistency score may be of doubtful value in detecting faking. Accordingly, it was not computed in this study.

Because of the large number of Spearman rhos computed, the .02 level was considered as indicating a significant degree of relationship or a significant difference.

Results

The hypothesis that there is an exclusive and representative pattern of attitudes or personality needs (as measured by the EPPS) among registered nurses which is shared by those seeking entrance into the profession, is not completely supported by the present study.

For each nursing group, the correlations with the other nursing groups are discussed first, followed by the correlations with non-nursing groups. These correlations are presented in Table 5.

Registered Nurses

The hypothesis that the EPPS patterns of groups of registered nurses would be significantly positively interrelated was supported. The comparison of the rankings for the registered nurses of General Hospitals I and II (RN GH I—RN GH II comparison) yielded a rho of .79 (a rho of .60 is significant at the .02 level for all comparisons made in this study). The registered nurse group from General Hospital I (RN GH I) also showed a significantly high degree of relationship with all the nursing groups (as can be seen in Table 5), the rhos ranging from .74 for the RN GH I—Seniors GH I comparison, to .95 for the RN GH I—Freshmen GH III comparison.

The registered nurse group from General Hospital II (RN GH II) showed a significant degree of relationship with the GH I freshman group (.61), with the combined freshman group (.70), with the GH IV Seniors (.73), and with the GH III freshman group (.76). However, the RN GH II group showed only a chance rela-

TABLE 5
Intercorrelations between EPPS Patterns of Nursing and Control Groups

	RN's GH I	RN's GH II	FGH I	FGH III	Combined F.	SGH I	SGH IV	A.W.	C.W.	HS G.
R.N.'s GH I	—					.74	.92	.54*	.92	—
R.N.'s GH II		.79	.90	.95	.93	.27*	.73	.79	.58*	—
Freshmen GH I		—	.61	.76	.70	.79	.94	—	.88	.89
Freshmen GH III			—	.92	—	.70	.91	—	.91	.84
Combined Freshmen				—	—	.76	.96	—	.89	.89
Seniors GH I						—	.77	.20*	.83	.89
Adult Women							—	.64	.82	.82
College Women								—	.45*	.45*
High School Girls									—	.89

*Not significant at .02 level.

tionship with the seniors of GH I (.27). These relationships were clearly not as great as those obtained between the RN GH I group and the student groups. (The correlation between the combined freshman group and the RN GH II group was significantly lower than that between the combined freshman group and the RN GH I group, $t = 2.43$, $p = .02$.) One might be prone to state that, since the interrelationships among the personnel of GH I were so high, the factor of setting or particularized training exerted a strong influence. However, the fact that the GH IV Seniors showed such a high degree of relationship with the RN GH I group (.92), with the Freshman GH I group (.94), and with the Senior GH I group (.77) indicates that setting or similarity of training is probably not an important factor in these EPPS pattern-relationships.

Further, although the EPPS pattern-relationships between RN and other nursing groups were high, relationships between RN and non-nursing groups were high as well.

For example, the RN's from GH I were found to share the need patterns of college women to a startling extent (.92) while those from GH II showed a significant EPPS pattern relationship with the Edwards norm-group of adult women (.79). On the other hand, the GH I RN group showed only a chance degree of relationship with the adult women group (.54) while the RN GH II group showed a chance degree of relationship with the group of college women (.58).

Thus, although registered nurses are similar to one another, there does not appear to be a specific and exclusive pattern of EPPS needs which characterizes RN's in diverse settings.

Freshmen

The two groups of freshman nursing students showed a significantly high relationship with all the other nursing groups in the study, the Spearman rhos ranging from .70 for the comparison with GH II RN's to .96 with the GH IV seniors. (See Table 5.) The freshman students, then, show great EPPS profile similarity with the other nursing groups.

However, they show a significant degree of pattern similarity to the group of high school girls (.89) and to college women (.89). As was found for the registered nurses, the EPPS pattern for the

combined freshman group does not serve to discriminate them from non-nursing controls.

Seniors

The EPPS patterns of the two groups of seniors showed a significantly high relationship with one another (.77). With the exception of the Seniors GH I—RN GH II comparison (.27) all the comparisons between each of the groups of seniors and other nursing groups were significantly high, ranging from .70 (Seniors GH I—Freshmen GH III) to .96 (Seniors GH IV—Combined Freshmen).

It is interesting to note that the GH IV seniors from South Carolina showed, overall, a higher degree of relationship with the St. Louis RN groups (RN GH I—.92, RN GH II—.73) than did the group of senior nursing students studying in St. Louis (RN GH I—.74, RN GH II—.27). (The differences between correlations were significant at the .02 and .001 levels, respectively.) This finding again casts doubt on the notion that need patterns are molded by specific settings.

Similarly, length of training does not appear to be influential as a determinant of specific occupational, personality-patterns as measured in this study since the registered nurse groups yielded EPPS patterns which tended to be more similar to those of freshman than to those of senior nursing students.

The EPPS patterns of GH I Seniors showed only a random relationship with the Edwards adult woman group ($\rho = .20$) but showed highly significant pattern relationships with both high school girls (.89) and college women (.83). The GH IV Seniors yielded patterns which were significantly similar to all three of these non-nursing groups (.64, .82, and .82 respectively). Thus the EPPS rankings of seniors show no exclusive "nursing" pattern.

The intercorrelations among the EPPS patterns of the three control groups reveal a high degree of similarity between high school girls and college women (.89). The correlations between high school girls and adult women and between college women and adult women however, do not indicate a significant degree of relationship (.45 in each case).

It is of interest to note that major discrepancies in EPPS rankings between the RN GH I and RN GH II groups were found for

three needs. For the nurses in GH I need heterosexuality placed 6 ranks higher than was the case for the GH II nurses, while needs order and endurance were placed 5 ranks lower by the GH I nurses. (See Table 4.) The differential ranking of needs heterosexuality and order determined to a great extent the correlations between patterns for each of the RN groups and the college and adult women samples. Both GH I nurses and college women ranked relatively low on need order (ranks 13 and 15 respectively), and relatively high on need heterosexuality (ranks 6 and 6 respectively), while GH II nurses and adult women ranked relatively high on need order (ranks 8 and 6 respectively) and relatively low on need heterosexuality (ranks 12 and 15 respectively). These same needs were among those which tended to attenuate the correlation between GH II nurses and the freshman groups and to maximize the relationship between GH I nurses and the freshman groups.

The GH II nurses and the GH I seniors ranked differently on this same core of needs.

Discussion

The core of needs found to influence most strongly the group interrelationships presented, has figured prominently in the studies of social desirability conducted by previous writers. Silverman (1957) employing male college students as subjects found, for example, that the MMPI scale of defensiveness, K, correlated significantly positively with EPPS needs deference and order, while Edwards' measure of social desirability (derived from other MMPI items) correlated significantly positively with need endurance. Klett and Tamkin (1957) employing neuropsychiatric patients noted, "The more deviant a person is (as defined by the MMPI) the less likely he is to consider endurance . . . as (a) socially desirable trait(s) and the more likely he is to judge heterosexuality as socially desirable" (p. 450). Krug and Moyer (1961) in a factor analytic study which included the EPPS scores of a group of male and female college freshmen defined a factor which they termed "compulsivity." This factor, which describes an individual who is "timid and somewhat anxious, minimizes his conflicts by behaving as much as possible according to established rules of procedure and keeps rather tight controls on himself at all times" (p. 294) showed high positive loadings on EPPS needs order, deference, endurance,

and abasement and high negative loadings on EPPS needs autonomy and heterosexuality.

It appears, then, that the GH II nurses may be a relatively conforming and compulsive group as compared with their counterparts (and with the students) at GH I. Since the two nursing groups do not differ significantly on demographic variables (see Table 3) the disparity between them may reflect differences in recruitment procedures or in work atmosphere between the two hospitals.

In general, then, the rank-correlation method, appears to demonstrate that, although there is a commonality of personal attitudes, as measured by the EPPS, among most nursing groups, these groups are not consistently discriminable from other groups of females who are not involved in nursing. The results of this study would tend to support the view that there is, in the case of nursing, a random relationship between personality-need patterns and choice of profession. Hospital or school "atmosphere," training, or selection may exert some influence on need patterns but this influence is, again, not a consistent one.

The virtual absence of an exclusive nursing pattern as defined in the present study may be explained in a number of ways. Accepting the present findings as veridical, they lend support to the view that personality needs may be secondary to extra-personality factors in determining selection or continuance in nursing. Or, perhaps, they may indicate that nursing is so broad a field that commonality of personality needs obtains only among subgroups of nursing personnel and may be diluted in considering a number of subgroups together.

An examination of the EPPS need definitions themselves may provide another possible explanation. These needs may themselves be too basic to provide discriminability of the type searched for in this paper. They may be measuring various facets and idiosyncracies of the "condition humaine" rather than discriminating suitability for or interest in nursing.

The nature of the population may provide still another explanation. It is well known that there is, presently, a dearth of professionally trained nurses. Thus, hospitals and nursing schools may have to limit their selectivity in an effort to produce a sufficient number of trained nurses. The fact that many nurses leave the pro-

fession, at least temporarily, within several years after entering it in order to marry and raise a family highlights one of the reasons why hospitals or schools cannot be overly selective in recruiting nurses or students. This fact also stresses the possibility that, for many women, entering a profession may simply be an expedient way of assuring themselves of financial support and independence prior to marriage. Habenstein and Christ (1955) have subsumed this nursing prototype under the term "utilizer." It implies the absence of total commitment to the profession and, *ergo*, the absence of a hypothetical personality pattern associated with membership in a professional group. In this instance, the elusive "nursing need pattern" might then be found among those nurses who remain active in and heavily committed to the field.

A final explanation for the absence of a "nursing need pattern" in this study is based on methodology. The use of all 15 raw scores of the EPPS yields ipsative data for each individual, since, for each protocol, the sum of the raw scores is the same. This results in the loss of one degree of freedom, since, if scores on 14 of the needs have been obtained, the score on the 15th is already determined. The need scores for each individual, then, are not completely independent of one another. This may have some effect on the rank-correlation procedure. Similarly, the use of need-rankings within each group, by obviating consideration of the number of subjects involved and the variance of the scores, takes no account of the reliability of the rankings obtained for each group. Certain of the findings may then be spurious.

Further research in this area might investigate the possibility that personality need-patterns are related to specific tasks within a vocation rather than to broad vocational designations such as "nursing."

Summary

1. The overall EPPS need patterns of two groups of registered nurses, two groups of freshman and two groups of senior nursing students were compared among themselves and with appropriate control groups by rank-correlational procedures.

2. The hypothesis that there is an exclusive and consistent pattern of personality needs among nurses which is shared by those entering or already in the nursing curriculum was tested. This hy-

pothesis was only partially supported by the data in that significantly high correlations were obtained between the rankings of the EPPS needs among the RN and nursing student groups. However, equally high correlations obtained between nursing groups and non-nursing controls. It appeared then that both nursing students and RN's were not consistently discriminable from control groups. Possible explanations for the failure to find an exclusive and consistent "nursing pattern" in this study were advanced.

REFERENCES

- Borislow, B. "The Edwards Personal Preference Schedule (EPPS) and Fakability." *Journal of Applied Psychology*, XLII (1958), 22-27.
- Edwards, A. L. *Edwards Personal Preference Schedule Manual Revised 1959*. New York: Psychological Corporation, 1959.
- Gynther, M. D. and Gertz, B. "Personality Characteristics of Student Nurses in South Carolina." *Journal of Social Psychology*, LVI (1962), 277-284.
- Habenstein, R. W. and Christ, E. A. *Professionalizer, Traditionalizer and Utilizer*. Columbia, Mo.: University of Missouri Press, 1955.
- Healey, Irene and Borg, W. R. "Personality Characteristics of Nursing School Students and Graduate Nurses." *Journal of Applied Psychology*, XXXV (1951), 275-280.
- Kibrick, Anne K. "Dropouts from Schools of Nursing: the Effect of Self and Role Perception." Unpublished doctoral dissertation, Harvard Graduate School of Education, 1958.
- Klett, C. J. "Performance of High School Students on the Edwards Personal Preference Schedule." *Journal of Consulting Psychology*, XXI (1957), 68-72.
- Klett, C. J. and Tamkin, A. S. "The Social Desirability Stereotype and Some Measures of Psychopathology." *Journal of Consulting Psychology*, XXI (1957), 450.
- Krug, R. E. and Moyer, K. E. "An Analysis of the F Scale: II. Relationship to Standardized Personality Inventories." *Journal of Social Psychology*, LIII (1961), 293-301.
- Kuder, G. F. *Examiner Manual for the Kuder Preference Record*. Chicago: Science Research Associates, 1951.
- Navran, L. and Stauffacher, J. C. "The Personality Structure of Psychiatric Nurses." *Nursing Research*, V (1957), 109-114.
- Reece, M. M. "Personality Characteristics and Success in a Nursing Program." *Nursing Research*, X (1961), 172-176.
- Redden, J. W. and Scales, E. "Nursing Education and Personality Characteristics." *Nursing Research*, X (1961), 215-218.
- Silverman, R. E. "The Edwards Personal Preference Schedule and Social Desirability." *Journal of Consulting Psychology*, XXI (1957), 402-404.

- Strong, E. K. *Manual for Vocational Interest Blank for Women*. Stanford: Stanford University Press, 1951.
- Williamson, Helen, M., Edmonston, W. E. Jr., and Stern, J. A. "The Use of the EPPS for the Identification of Personal Role Attributes Desirable in Nursing." *Journal of Health and Human Behavior*, in press.
- Zuckerman, M. "The Validity of the Edwards Personal Preference Schedule in the Measurement of Dependency-Rebelliousness." *Journal of Clinical Psychology*, XIV (1958), 379-382.



FACTOR ANALYSIS OF RANKED EDUCATIONAL OBJECTIVES: AN APPROACH TO VALUE ORIENTATION¹

FRED W. OHNMACHT
University of Maine

It is safe to assume that qualified personnel in the field of education differ with respect to the relative value or importance which they are willing to ascribe to elements subsumed under the rubric of general objectives of education. Discrepancies of this sort probably become reasonably clear where there is active debate with a concomitant forthright expression of opinion. Frequently, however, discussions which appear to be activity of this sort are couched in polite language embodying sufficient flexibility so as not to endanger the feelings of those participating. If this is the case, or where little or no interaction takes place, the differing points of view do not become crystallized and remain latent.

Methods of an experimental nature which produce indications of such latent controversies or divergent value orientations should aid materially in the description of and differentiation among groups of individuals involved in the educational enterprise. One possible approach to this problem is the factor analysis of ranked educational objectives.

Purpose of Study

The purpose of the present study is two-fold: (1) to investigate the utility of factor analysis of ranked educational objectives for

¹Based upon research supported by the College of Education Team Teaching Project which is funded by the Ford Foundation. The author wishes to acknowledge the computational assistance of Mr. Russell Altenberger, University of Maine Computing Center and Dr. Charles Grant who helped in the data collection.

the purpose of identifying value orientation in terms of the relative importance attributed to educational objectives as contained in a selected list; and (2) to identify the value orientation, within the context of a sample of objectives, of a College of Education faculty.

Method and Procedures

Initially it was necessary to assemble a list of general objectives of education. It was decided to use the ten classifications of general educational objectives developed by the Commission on the Relation of School and College of the Progressive Education Association (Smith, 1942). The ten statements were listed with their order being determined using a table of random numbers. The directions for the task and statements in the order they were given is presented here.

Please rank the following statements of educational objectives in order from 1 to 10. Assign number 1 to the objective you feel is most important and number 10 to the objective you feel is least important.

- _____ The acquisition of important information.
- _____ The development of effective methods of thinking.
- _____ The development of increased appreciation of music, art, literature, and other aesthetic experiences.
- _____ The development of social sensitivity.
- _____ The acquisition of a wide range of significant interests.
- _____ The development of better personal-social adjustment.
- _____ The development of physical health.
- _____ The inculcation of social attitudes.
- _____ The cultivation of useful work habits and study skills.
- _____ The development of a consistent philosophy of life.

Twenty professors associated with the College of Education at the University of Maine were asked to perform the task indicated above. Fifteen of these professors represent three teaching teams of five each. Each team is responsible for one of three core offerings in the professional sequence for undergraduate students in the College of Education. Five of the respondents represent professors involved in the administration of the college or at least are not identified with teaching the basic three-course sequence.

After obtaining the rankings from the 20 professors, it is of interest to inquire whether the rankings are random across the respondents, or whether there is a systematic agreement among groups of professors. To answer this question an obverse factor analysis was undertaken using each professor as a variable.

The matrix of intercorrelations containing the relationships of each person's rankings with every other set was computed. This matrix then represents the intercorrelations among individuals rather than the relationships among a set of tests. The correlations were computed using the rank-difference method (Spearman rho).

The matrix was factored, with unities in the diagonal, using the principal-axes method. All of the principal components whose latent root exceeded one were retained for the purpose of rotation. This procedure follows the recommendations of Kaiser (1960a, 1960b) who found that the generalized Kuder-Richardson reliability of a principal component will be positive when its associated latent root is greater than one. Guttman (1954) demonstrated that the number of common factors in a domain must be at least as large as the number of latent roots which are greater than one when unities are placed in the diagonal. Using the above criterion, six factors were extracted and were then analytically rotated using the Normal Vari-max Solution (Kaiser, 1958). Individuals with a high loading on a given factor should have systematically similar rankings.

To assist in the interpretations of the factors, the rankings of individuals having loadings in excess of .70 on a factor were examined. The median ranking for each statement was calculated and the factor interpreted on the basis of those statements receiving high and low composite ranks.

Findings

Table 1 presents the matrix of intercorrelations among the rankings of the 20 professors who comprise the sample. An examination of the matrix indicates that systematic agreement does exist among certain sub-groups of the total sample. For example individual 4 correlates .60 or better with eight other individuals in the sample and one of these correlations exceeded .90. This systematic agreement is brought out in a more precise form in the factor analysis to be discussed.

Table 2 presents the loadings for each individual on the six principal components whose latent roots exceed unity. It should be remembered that unities rather than communality estimates were placed in the diagonal of the correlation matrix for the purpose of factoring. The resulting analysis is more properly called com-

TABLE 1

*Matrix of Intercorrelations among Rankings of Educational Objectives by Twenty Personnel Associated with Teacher Education**

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	23	28	14	82	25	-01	25	26	26	-17	14	01	47	34	-25	-12	-34	43	29
2		—	41	76	67	52	76	95	83	52	40	58	28	37	66	20	51	49	20	19
3			—	24	53	61	34	48	36	29	31	19	13	29	81	14	37	49	11	02
4				—	54	57	92	58	78	63	29	76	60	36	65	55	80	48	32	-31
5					—	60	40	69	61	53	11	54	26	66	63	02	32	10	52	26
6						—	53	54	67	34	-16	66	32	65	65	35	80	28	49	12
7							—	61	74	66	41	65	58	31	70	66	72	64	04	-20
8								—	70	52	37	57	22	47	60	13	40	51	17	43
9									—	24	05	45	17	34	66	19	65	17	43	02
10										—	60	77	86	66	40	72	42	65	01	13
11											—	18	55	10	22	49	10	72	-46	07
12												—	75	69	40	60	76	53	41	07
13													—	60	18	90	60	59	07	-13
14														—	18	48	40	13	54	51
15															—	19	57	52	01	-20
16																—	63	61	07	-18
17																	—	52	31	-31
18																		—	-46	11
19																			—	10
20																				—

*Decimals omitted.

TABLE 2

Orthogonal Factor Loadings Extracted by the Principal Axes Solution

Person	Factor					
	I	II	III	IV	V	VI
1	26	73	15	26	27	46
2	82	15	-33	17	-38	06
3	55	16	-40	20	61	-22
4	88	-09	-09	-33	-19	27
5	71	59	03	22	12	26
6	76	34	01	-26	20	-42
7	86	-26	-20	-17	-17	13
8	77	22	-26	37	-32	-12
9	74	34	-32	-25	-31	02
10	78	-29	34	31	08	23
11	40	-60	-11	55	00	19
12	84	-05	36	-12	-11	-08
13	68	-48	51	-02	12	14
14	65	30	62	19	08	-16
15	74	12	-56	-03	33	07
16	61	-59	40	-16	10	-06
17	79	-15	-01	-51	05	-24
18	63	-62	-21	31	06	-25
19	30	66	40	-40	-06	-03
20	06	37	26	72	-27	-42
Latent Root	9.16	3.41	2.16	2.10	1.15	1.06
% of total variance	45.77	17.06	10.81	10.50	5.79	5.32

ponent analysis which employs a somewhat different mathematical model than does factor analysis. Kaiser has pointed out, however, that when components whose latent roots exceed one are rotated they may be interpreted substantively in the same way as factors derived under the model which uses estimated communalities in the diagonal. Thus there are six discernible types of performance underlying the ranks assigned by the 20 professors. It is these types of performance which will be called relative value orientation.

Table 3 presents the loadings of each individual on the six normal varimax factors. Only those loadings which exceeded .40 are listed to clarify the basic pattern.

The varimax solution places the restriction of orthogonality upon the rotation and tends to produce a solution which meets the criteria of simple structure when the restriction imposed adequately reflects the nature of the data. An examination of Table 3 reveals that nine individuals have a complexity of one and that

TABLE 3

Factor Loadings Derived from Normal Varimax Solution

Person	Factor					
	I	II	III	IV	V	VI
1						94
2			92			
3					93	
4	53		78			
5			48			74
6		50	42		68	
7	52		75			
8			79	46		
9			88			
10	82					
11	42	-80				
12	76		45			
13	98					
14	64			52		
15			57		73	
16	92					
17	58		51		40	
18	51	-58			41	
19		82				
20				98		
Hyperplane count ± 15	7	9	5	9	9	7

every factor, with the exception of number three, has more than six loadings within a $\pm .15$ band. It would seem that an orthogonal frame is a reasonable one from which to attempt an interpretation. Several factors are represented by individuals who ranked a given objective in a similar relative position, but the differences in their rankings are compelling enough to suggest a low relationship between the factors thus represented.

Interpretation of Factors

Each of the six factors, with the exception of number three, have at least one individual with a loading which exceeds .90. In the cases where the factors had a large number of high loadings (Factors I and III) the rankings of all individuals with a loading in excess of .70 were assembled and a median rank for each statement computed. Since strong agreement for certain statements was apparent the computations will not be presented here, but will be summarized below. Factors IV and VI appeared to be rather specific fac-

tors which could be characterized by the rankings of but one or two individuals. For Factor V an individual with a loading of .63 was included in the analysis for interpretive purposes for that factor. Factor II had only two individuals with a loading in excess of .70, one representing each pole, but upon examination of their rankings, in conjunction with two other individuals having moderately high (+50, -52) loadings on this factor, the nature of the polarity became quite clear. A brief description of the characteristics of the rankings of individuals having a high loading on each factor with a tentative label for each factor is presented here. The interpretations should be thought of more as hypotheses for further investigation rather than as confirmed constructs.

Factor I—Process

This factor was most strongly represented by individuals ranking effective methods of thinking and the cultivation of useful work habits and study skills as being most important. The development of physical health and the inculcation of social attitudes were the statements ranked at the very bottom of the hierarchy. Although the agreement was not as pronounced as for the above statements, there was a tendency on the part of persons having high loadings to rank the acquisition of a wide range of interests at the upper end of the hierarchy with a concomitant low ranking for the development of a consistent philosophy of life.

Factor II—Self-sufficiency

This factor was bi-polar in form. Persons having loadings defining the two ends of the continuum agreed that effective thinking should be near the top in terms of relative importance. The most striking feature of the rankings of the polar opposites was a reversal of rankings for a consistent philosophy of life and the cultivation of useful work habits and study skills, the positive pole being characterized by high rankings for these statements (in the top three) and the negative pole being characterized by low rankings (in the bottom three). Individuals representing the positive pole tended to rank personal-social adjustment low as opposed to relatively higher rankings by those representing the negative pole. The poles are further characterized by reversals for aesthetic appreciations and broad interests with persons at the positive pole

ranking these in relatively lower positions than those at the negative end.

The positive pole may represent persons who see effective thinking as important within the context of the development of a consistent personal philosophy. The negative pole may represent persons who view effective thinking as important within a context of interactions with others reflected by the relatively high rankings given by these persons to the statement concerning personal-social adjustment and broad interests and aesthetic appreciations. This does not imply a rejection of any of these statements, but simply a relative preference. The general pattern described may be characterized as an emphasis upon the abilities needed to be autonomous or inner directed as opposed to a development of abilities which optimize adjustment with others.

Factor III—Content

This factor seems quite clear in that persons high on this factor consistently rank effective thinking and the acquisition of important information as either one or two in the hierarchy. Factor III is also represented by persons who uniformly ranked the development of physical health and increased aesthetic appreciations as being least important of the objectives listed. These persons also tended to rank the development of a consistent philosophy as being among the more important objectives while ranking the cultivation of useful work habits and study skills and inculcation of social attitudes low. In contrast to individuals representing Factor I, the remaining objectives concerning social sensitivity and personal-social adjustment were ranked relatively low by persons with high loadings in Factor III. In both cases the rankings were in the middle range (4 to 7).

Factor IV—Controlled Personal—Social Adjustment (Specific)

Factor IV appears to be a specific factor defined by one individual with a loading of .98. Two other individuals have only moderately high loadings on this factor. The outstanding characteristic of the ranking of individual 20, who essentially represents this factor is the priority given to the inculcation of social attitudes in conjunction with high rankings for the development of social sensitivity and personal-social adjustment. Aesthetic appreciations, a

wide range of interest, and the development of physical health received low rankings.

Factor V—Reflective Awareness

This factor is defined by two people with loadings in excess of .70 with several others having loadings in excess of .40. The development of effective methods of thinking and of social sensitivity received high priority from individuals representing this factor. The development of physical health and effective work habits and study skills were ranked at the bottom. The development of a consistent philosophy of life also received high rankings across individuals with a less pronounced tendency to also rank aesthetic appreciation high. A wide variety of interests tended to receive relatively low rankings. The factor seems to reflect an emphasis upon effective thinking within a context of reflective thought as underscored by the relatively high rankings for the social sensitivity and consistent philosophy statements. It is interesting to further note that the individuals who define this factor gave the personal-social adjustment statement a relatively low ranking.

Factor VI—Adjustment

Factor VI appears to be a relatively specific factor which is characterized by the rankings of two respondents. The rankings assigned to consistent philosophy, personal-social adjustment, and social sensitivity were high while aesthetic appreciations, acquisition of important information and physical health were ranked low. The cultivation of useful work habits and study skills was ranked seventh by both individuals representing this factor which is also a relatively low position in the hierarchy.

The foregoing indicates that Factors I and III are the most clearly established in terms of the number of individuals having a loading in excess of .40 on each. Factors IV and VI tend to be specific to several individuals and represent variations of relative emphasis on social adjustment. Factor II is clearly bi-polar in form and the reversals of certain rankings are clear. Its interpretation and label as presented here are quite tentative. Factor V, although not as clearly established as I and III, appears to represent a distinguishable point of view.

Conclusions and Implications

1. The factor analysis of ranked educational objectives has been demonstrated to be an effective method for differentiating among a group of individuals in terms of the relative importance they attribute to the elements of a set of objectives. Relative importance is assumed to represent value.

2. In terms of the present study, six factors representing systematic agreement among various sub-groups of the sample were extracted. These factors were tentatively identified as:

- I. Process
- II. Self-sufficiency
- III. Content
- IV. Controlled Personal-Social Adjustment
- V. Reflective Awareness
- VI. Adjustment.

The above is not meant to imply that individuals who are most adequately represented by a given factor do not value those objectives which are given low rankings. The factors reflect a relative emphasis in terms of the operations required of each respondent.

The method would seem to have utility with any set of objectives, of reasonable length, as a heuristic prelude to the posing of hypotheses with regard to interpersonal perception, interaction patterns among groups, teacher behavior in the classroom, and the like. We could raise the question of student value orientation in terms of course objectives and the orientation of a team of teachers which has responsibility for the course. In what other ways do individuals, representing factors uncovered utilizing the proposed method, differ? In essence the factors uncovered utilizing the method can be thought of as hypothetical constructs which can be introduced into a network of other such constructs represented by test scores and other observables.

REFERENCES

- Guttman, Louis. "Some Necessary Conditions for Commonfactor Analysis." *Psychometrika*, XIX (1954), 149-161.
- Kaiser, Henry F. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika*, XXIII (1958), 187-200.

- Kaiser, Henry F. "The Application of Electronic Computers to Factor Analysis." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XX (1960), 141-151. (a)
- Kaiser, Henry F. "Comments on Communalities and the Number of Factors." A paper read at an informal conference at Washington University, St. Louis, May 14, 1960. (b)
- Smith, Eugene R. et al, Progressive Education Association, Commission on the Relation of School and College, *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. p. 18.

READING DIFFICULTY OF PHYSICS AND CHEMISTRY TEXTBOOKS

MILTON D. JACOBSON
University of Virginia

TEXTBOOK selection is an important but difficult task. The primary reason for the difficulty is the lack of a foolproof method for evaluating textbooks. Methods now in use lack objectivity and are too often based on friendship with the salesman or on the attractiveness of a cover (Mallinson, 1958).

The factors for evaluating textbooks are manifold. However, among the most important is the factor of reading difficulty. Major (1955) has shown experimentally that decreasing reading difficulty increases the learning which takes place for students reading biology material at both the high school and college levels. Marshall (1956) assumed that this would also be true for physics and chemistry materials but concluded that the Flesch formula was not appropriate for determining or for modifying reading difficulty in these areas. He recommended that educators discontinue using this and other reading formulae for predicting the difficulty of physics and chemistry materials unless their use has been validated experimentally. To date this validation has not been done satisfactorily for any formula in the literature.

This study had two objectives. The first was to determine the relative reading difficulty of physics and chemistry textbooks in use in Minnesota public schools. This objective will be considered in Part I of this study. The second objective, which will be considered in Part II, was to compute regression equations which would be valid for predicting the reading difficulty of physics and chemistry textbooks.

Part I

Reading Difficulty

Part I of the study was concerned with testing six null hypotheses concerning differences in reading difficulty. These are listed below and will be dealt with in detail in the section on analysis.

1. There is no between school effect.
2. There is no between unit effect.
3. There is no between student effect.
4. There is no between textbook effect.
5. There is no school by unit interaction effect.
6. There is no student by school interaction effect.

Design and Procedure

Reading difficulty was determined by means of the *Underlining Test*. This test was first reported by Kyte (1928) was used extensively by Curtis (1938) and was applied to a large number of physics and chemistry materials by Warriner (1951).

The reliability of the *Underlining Test* was established by the test-retest procedure in a preliminary study. Product-moment reliability correlation coefficients were 0.95 for physics and 0.85 for chemistry. The time interval between the test and retest was one day for physics and one week for chemistry.

In applying this test, students enrolled in chemistry or physics classes in randomly selected Minnesota public high schools, read sample passages from chemistry or physics books, respectively, and underlined every word which they did not understand. The number of underlined words gave the measure of the reading difficulty of the passage.

Rank-difference correlation coefficients were calculated for a student's rank on a vocabulary test constructed by the author and on the number of words underlined. The majority of these coefficients were negative and showed that students with larger vocabularies underlined fewer words. This was in agreement with the findings of Curtis and Warriner, and it was concluded that it was valid to use the *Underlining Test* to determine reading difficulty.

A previous survey of Minnesota public high schools was made in

the spring of 1959 to determine the books which were in use, to rank them according to popularity, and to obtain information necessary for defining a population of schools from which science classes could be randomly selected for inclusion in the study. The school population was limited to those which would offer physics and/or chemistry in the fall of 1959 and would have an expected enrollment of at least 16 students. Chemistry classes from 10 schools and physics classes from 12 schools were used in the study.

Since the precision of the regression equations would be proportional to the variability in the difficulty levels of the reading materials, the variability was increased by including college textbooks in the study which were in use at the University of Minnesota in freshman and sophomore classes. The differences between any two books would not be affected by the inclusions mentioned, and comparisons of these differences could be determined by multiple F analysis.

In physics 13 books from the preliminary survey were selected and were supplemented with two college physics textbooks and one high school physical science textbook.

For chemistry the 13 textbooks from the survey were supplemented with three college textbooks.

Thus, the passages to be underlined by the students were selected from 16 textbooks in physics and from 16 textbooks in chemistry.

The textbooks were separated into 10 units according to their content (e.g., sound and heat in physics and gases and solutions in chemistry) with each unit comprising about 10 percent of most books. Samples of about 200 words of continuous material were randomly selected from each unit. These were reproduced by the multilith process and assembled into booklets for each student. The booklets were assembled for each school according to a 10 by 16 incomplete Latin (Youden) square design in which books were assigned to treatments, units to columns, and students to rows of the square. In every school each of 16 students received one booklet containing 10 samples, one from each of the 10 units. Each sample was, however, taken from a different book. In each school every sample was read once, and no two students read the same sample from any one book.

Analysis

The 6 null hypotheses referred to in the beginning of this paper were tested by the analysis of variance technique appropriate for this design. A significance level of 5 percent was selected at the time of the design of the investigation for determining the acceptance or rejection of the hypotheses under examination. The results of the analysis of variance are summarized in Table 1 for chemistry and Table 2 for physics.

TABLE 1
*Analysis of Variance of the Underlining Scores
for Reading Samples Taken from Chemistry Books*

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F	Hypothesis
Columns					
Between Schools	9	2336.926	259.6584	10.2316**	Rejected
Between Units	9	1133.563	125.951	4.962**	Rejected
School by Unit	81	5900.318	72.843	2.870**	Rejected
Rows					
Between Students	15	1389.570	92.638	3.6503**	Rejected
Students by School	135	12,711.624	94.1602	3.7103**	Rejected
Treatments (Between Books)	15	4752.559	316.837	12.4846**	Rejected
Error (by subtraction)	1335	33,879.935	25.3782		
Total	1599	62,104.495			

**Significant at the 1 percent level.

Since the books were different in reading difficulty, Scheffé's multiple F test was utilized to determine which books were more difficult than others at the 5 percent level of significance.

In Table 3 the difficulty and popularity rankings of the chemistry books are shown along with the results of Scheffé's test. The books are identified by means of book numbers previously assigned and by their classification as high school or college textbooks.

Table 4 presents data for physics which is analogous to that in Table 3 for chemistry.

Findings

The hypothesis of no difference between the reading difficulty of the different textbooks was rejected at the 1 percent level of significance for both the chemistry and the physics studies.

TABLE 2

*Analysis of Variance of the Underlining Scores for
Reading Samples Taken from Physics Books*

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F	Hypothesis
Columns					
Between Schools	11	1259.2800	114.48	6.451**	Rejected
Between Units	9	345.156	38.3507	2.161*	Rejected
School by Unit	99	5364.038	54.182	3.053**	Rejected
Rows					
Between Students	15	1064.487	70.9658	3.999**	Rejected
Student by School	165	10,512.545	63.7124	3.590**	Rejected
Treatments					
Between Books	15	6264.6588	417.6439	23.535**	Rejected
Error					
(By subtraction)	1605	28,482.247	17.7459		
Total	1919	53,292.412			

**Significant at the 1 percent level.

*Significant at the 5 percent level.

TABLE 3

Difficulty Differences versus Popularity of Chemistry Books

Book Number	Classification	Popularity Ranking	Difficulty Ranking	Mean Underlining Score, T_1
14	College	Not determined	1 ^a	10.627
4	College	Not determined	2 ^a	8.943
6	College	Not determined	3 ^b	6.420
12	H.S.	9	4	6.276
7	H.S.	12	5	5.882
10	H.S.	9	6	5.534
2	H.S.	1	7	5.281
11	H.S.	12	8	4.929
5	H.S.	4	9	4.909
8	H.S.	5	10	4.719
9	H.S.	3	11	4.622
3	H.S.	6	12	4.473
1	H.S.	2	12	4.473
15	H.S.	11	14	4.224
16	H.S.	7	15	4.175
13	H.S.	8	16	3.981

^aRank numbers 1 and 2 are both significantly more difficult than any of the other 14 books.

^bRank number 3 is significantly more difficult than rank number 16.

TABLE 4

Difficulty Differences versus Popularity of Physics Books

Book Number	Classification	Popularity Ranking	Difficulty Ranking	Mean Underlining Score, T_1
4	H.S.	10	1 ^a	9.227
11	H.S.	8	2 ^b	5.383
6	College	Not determined	3 ^c	5.017
1	College	Not determined	4 ^c	5.015
12	H.S.	8	5 ^d	4.194
16	H.S.	5	6 ^d	4.010
8	H.S.	4	7 ^e	3.759
10	H.S.	10	8 ^e	3.534
15	H.S.	1	9 ^f	3.427
7	H.S.	10	10 ^g	3.227
2	H.S.	8	11	2.811
3	H.S.	2	12	2.374
14	H.S.	6	13	2.058
5	H.S.	6	14	1.963
13	Phys. Sci.	Not determined		1.709
9	Phys. Sci.	13	16	1.561

^aRank number 1 was more difficult than rank numbers 2-16.^bRank number 2 was more difficult than rank numbers 7-16.^cRank numbers 3 and 4 were more difficult than rank numbers 8-16.^dRank numbers 5 and 6 were more difficult than rank numbers 12-16.^eRank numbers 7 and 8 were more difficult than rank numbers 13-16.^fRank number 9 was more difficult than rank numbers 14-16.^gRank number 10 was more difficult than rank numbers 15-16.

Inspection of Tables 3 and 4 shows that the differences between physics books were more pronounced than were those between chemistry books. The two most difficult physics books were in use in high school, while in chemistry two college textbooks were more difficult than any of the other books considered. The most difficult physics book had a low popularity rank and was significantly more difficult than the other 15 books considered. The second most difficult physics book was significantly more difficult than 10 of the books considered. This book was the third most widely used book in state high schools.

To determine if the most readable books were being more frequently used in public schools, rank-difference correlation coefficients were calculated comparing the reading difficulty rank of a book with its popularity rank which had been determined in the preliminary survey. The data in Table 3 were used to obtain the correlation coefficients for the chemistry books and that from

Table 4 for the physics books. These coefficients were -0.434 for chemistry and $+0.29$ for physics. Neither coefficient was significant, and it was concluded that reading difficulty was not a factor in the evaluation of textbooks by those responsible for these selections.

The hypothesis that there is no difference between the underlining scores on different units was rejected at the 1 percent level of significance for the chemistry study and at the 5 percent level for the physics study. It was concluded that there were significant differences in the reading difficulty among units from both the chemistry and the physics books.

Tables 1 and 2 also show that there were differences between schools, and also differences between students and that there were school by unit and student by school interactions. These were all significant at the 1 percent level.

Part II

Covariance and Regression Analysis

Covariance and regression analyses were made to obtain equations for predicting the reading difficulty of physics and chemistry passages by means of independent variables determined from the passages.

The data for the analysis of covariance consisted of six measurements.

1. The average underlining score (Y) for each reading sample which was obtained by averaging, for each sample, the number of words underlined by all of the students reading the sample and was the dependent variable.

The following measurements are independent variables.

2. The average number of words per independent clause in each sample passage, (X_1).
3. The concentration of simplified mathematical terms (X_2) in each sample passage.
4. The concentration of words in each sample passage which are above the 6,000 word level in Thorndike's 20,000 word list, (X_3).
5. The concentration of words in each sample passage which are not included in Powers word list of 1828 essential scientific terms, (X_4).

6. The ratio of the average number of syllables per word to the total number of words in each passage, (X_5).

Design for Covariance Analysis

The design chosen gave a two-way variance breakdown according to the level of reading difficulty of each unit and of each book to obtain the sums of squares and of products of the dependent and the five independent variables. The sums of squares and of products which were computed in the sources of variation attributable to error and to the total in this table became the terms of two 6×6 matrices. These matrices were inverted to obtain multiple regression coefficients which were tested for significance. This was done separately for the chemistry and physics studies.

The analysis of covariance and the regression analysis were accomplished by using an electronic computer which expedited the process and increased the accuracy of the calculations.

Findings

In these findings y will designate the reading difficulty and the X 's with subscripts will represent the independent variables described in the beginning of Part II.

The equations and their multiple correlation coefficients, R , with the dependent variable were, for chemistry:

- a) book-unit effect extricated: $R = 0.581$,

$$y = -.0004 + 26.8281X_2 + 39.1307X_4 + 1.5574X_5;$$

- b) no book-unit effect: $R = 0.702$

$$y = -0.0013 + 29.7059X_2 + 21.1119X_3 + 35.0029X_4;$$

and for physics:

- a) book-unit effect extricated: $R = 0.521$,

$$y = -0.0229 + 0.0736X_1 + 13.9025X_2 + 20.2393X_3;$$

- b) no book-unit effect: $R = 0.674$,

$$y = 0.003 + 0.1706X_1 + 13.7231X_2 - 43.7262X_3 - 2.3577X_5.$$

Recommendations

It is recommended that: (1) reading difficulty be an important criterion for selecting physics and chemistry textbooks; (2) the equations developed be used to determine the reading difficulty of physics and chemistry material; and (3) future reading difficulty studies be conducted separately for physics and chemistry and

further consideration be given to the variables appearing in the four regression equations.

REFERENCES

- Curtis, Francis D. *Investigations of Vocabulary in Textbooks of Science for Secondary Schools*. Boston: Ginn and Company, 1938. Pp. 21-36.
- Kyte, George C. "Experimentation in the Development of a Book to Meet Educational Needs." *Education Administration and Supervision*, XIV (1928), 86-100.
- Major, Alexander G. "Readability of College General Biology Textbooks and the Probable Effects of Readability Elements on Comprehension." Unpublished Ph.D. thesis, Syracuse University, 1955.
- Mallinson, G. G. "Textbook and Reading Difficulty in Science Teaching." *Science Teacher*, XXV (1958), 474-5.
- Marshall, James S. "The Relationship between Readability and Comprehension of High School Physics Textbooks." Unpublished Ph.D. thesis, Syracuse University, 1956.
- Warriner, David A. "An Investigation of the Effect of Certain Psychosemantic Factors of the Level of Reading Comprehension Difficulty in High School Chemistry and Physics." Unpublished Ph.D. thesis, Cornell University, 1951.

ENGINEERING FRESHMAN NORMS FOR THE D. A. T. MECHANICAL REASONING AND SPACE RELATIONS TESTS UTILIZING FIFTEEN-MINUTE TIME LIMITS

CHARLES W. JONES AND DAN McMILLEN

Iowa State University

VOCATIONAL counselors are constantly in search of tests and other measuring devices which will help predict the probable success of their counselees in various fields of endeavor. It was this type of consideration that motivated the investigators to conduct the following study. The authors feel that the abilities of the engineer must be quite versatile and wide in scope and also assume that practically all engineers, whether in the course of their training or in their later occupational function, will, of necessity, have recourse to use the abilities described as spatial and mechanical. The engineer finds it necessary to deal with mechanical objects contained in space. His abilities to understand relationships of objects in space and to make graphical presentations of his concepts and ideas are considered to be important in determining his academic success and possibly his occupational success. These are some of the reasons the authors selected a spatial and a mechanical test for consideration.

A follow-up study, conducted by Bennett, Seashore, and Wesman (1952), and reported in the *Manual for the Differential Aptitude Tests*, included results on the entire battery for 70 engineering students. It was found that these students were "Superior in Numerical Ability, and outstanding as compared with all other groups in Space Relations and Mechanical Reasoning."

The following information relevant to the usefulness of the Dif-

ferential Aptitude Tests is also contained in the D. A. T. manual:

At the Institute of Technology of the University of Minnesota, the tests are not quite so suitable in level of difficulty or spread of scores. Apparently the freshmen at the Institute are so selected as to constitute a group noticeably superior to the twelfth graders and to the freshmen at the other three colleges (Kansas State Teachers College, Oswego State Teachers College, and Allegheny College). The Minnesota students are also more homogeneous in the abilities measured by several of the tests. The relatively small spread of scores perhaps accounts for some of the validity coefficients. At the same time, it makes more noteworthy the coefficients (such as those for the Verbal Reasoning and Numerical Ability tests) which demonstrate appreciable predictive capacity.

The Instruments

The Differential Aptitude Test Battery was developed by Bennett, Seashore, and Wesman for general use at the secondary school level, and it contains tests measuring seven factors: Mechanical Reasoning, Spatial Relations, Verbal Reasoning, Numerical Ability, Abstract Reasoning, Clerical Speed of Accuracy, Language Usage. The tests used in this study were the Mechanical Reasoning Form B and the Spatial Relations Tests, Form B. The directions for the D. A. T. tests could be read and understood without difficulty by most college freshmen in a very brief time, and this lent credence to the use of the tests in the manner intended by the investigators.

The Sample

The following curricula of the College of Engineering at Iowa State University are represented in the sample.

Curriculum	N
Aerospace Engineering	83
Agricultural Engineering	36
Architecture and Architectural Engineering	104
Ceramic Engineering	6
Chemical Engineering	82
Civil Engineering	80
Electrical Engineering	164
Engineering Operations	2

Engineering Science	1
Industrial Engineering	31
Mechanical Engineering	78
Total	667

Procedure

The space and mechanical test of the D. A. T. are normally timed at 30 minutes each. The decision was made to administer these tests to groups of freshman engineering students using a 15 minute time limit for several reasons. Most important, perhaps, is the fact that on these two tests engineering students tend to make raw scores which are high in general, and the resulting skewed distribution discriminated well at the lower end of the continuum but provided little basis for differential at the upper end. It was believed by the investigators that the resulting distributions from the fifteen minute time limitations would increase the predictive value of the instruments and would thus contribute to counseling usability at the college as well as, perhaps, the 12th grade level. Members of the Student Counseling Service at Iowa State University find the published norms of the two tests inadequate when engineering students are administered the tests for individual counseling purposes. It may be well to add that students in the engineering college report speed to be a factor of achievement in such classes as those concerned with graphics, sliderule, and mathematics. Another reason for cutting the time to 15 minutes involves strictly time expediency. Each group of students traditionally meet in an engineering orientation lecture session for a period of 50 minutes. Such time allowed the administration of both tests at one sitting.

With regards to the administration of the tests, it should also be noted that the mechanical and spatial tests were dispersed in an alternating pattern at the time of testing. Both tests were handed to the students at the same time. They were fastened together in such a way as to present initially an alternate test to succeeding students in a row. Every other student was confronted with the Mechanical test, and alternate students took the Spatial test during the first test interval. The second test interval subsequently found, again, alternating tests along each row of students. It was hoped that copying would be minimized by using this approach, and also a correction for possible fatigue factor would be operating.

In the introduction to the testing process, it was mentioned to

the students that the tests were usually timed for 30 minutes each, but that in this case a 15 minute interval would suffice. These remarks may have had the effect of accelerating some of the students in their approach to the testing situation. It is suggested that anyone who plans to utilize these norms make essentially the same remarks to the examinees in order to control for this effect. The examinees were also responsible for reading and understanding the directions. No pointers were given and no assistance offered.

The testing environment under which the engineering freshmen were tested seemed to be quite adequate, and the students appeared to be motivated to pursue the test items with enthusiasm.

Subsequent data concerning first quarter engineering graphics, second quarter cumulative grade point average, and first year cumulative grade point average were made available to the investigators. The results of the Otis Quick Scoring Mental Ability Test, Gamma form, which is administered to all entering freshmen, were also considered.

The D. A. T. and Otis MA scores were translated by utilizing McCalls Normalizing Method to T scores to facilitate the handling of combinations of data. Product-moment correlation coefficients are presented for relevant combinations in Table 3.

Results

In spite of the fact that the time periods for the two tests were cut in half, it is noteworthy to observe that three percent of the students completed all items on the spatial test, and that nineteen percent of the students completed all items on the mechanical test. The speed factor which was introduced into these tests resulted in a much larger range of scores for each of the tests. The relatively small spread, which occurred in the study made on Minnesota Institute of Technology students, would most likely also have been found in this study if the 30 minute time limit had been followed for each of the tests.

Table 1 represents the resulting norms for freshman engineering students at Iowa State University using the Space Relations test from the D. A. T. with the 15 minute time limit. Scoring formula applied: $R-W$.

Correlation coefficients contained in Table 3 indicate relatively low positive relations existing between the scores achieved by the

TABLE 1

*Percentile Norms for the D. A. T. Space Relations Test
with Fifteen-Minute Time Limit
Norm Group—Iowa State University Freshman Engineering Students*

Raw Score	Percentile	Raw Score	Percentile	Raw Score	Percentile
83+	99	57-58	65	36-39	20
79-82	98	54-56	60	30-35	15
77-78	97	52-53	55	27-29	10
73-76	95	50-51	50	22-26	5
70-72	90	47-49	45	20-21	4
66-69	85	45-46	40	17-19	3
63-65	80	44	35	15-16	2
61-62	75	42-43	30	9-14	1
59-60	70	40-41	25	-20-8	0
Mean = 50.12		Range = -20 to 96		SD = 17.10	N = 667

TABLE 2

*Percentile Norms for the D. A. T. Mechanical Reasoning Test
with Fifteen-Minute Time Limit
Norm Group—Iowa State University Freshman Engineering Students*

Raw Score	Percentile	Raw Score	Percentile	Raw Score	Percentile
66+	100	51-52	65	40	20
65	99	49-50	60	37-39	15
63-64	98	48	55	34-36	10
61-62	95	47	50	32-33	5
59-60	90	46	45	31	4
57-58	85	45	40	30	3
55-56	80	44	35	27-29	2
54	75	43	30	15-26	1
53	70	41-42	25	0-14	0
Mean = 47.26		Range = 14 to 66		SD = 9.11	N = 667

Scoring formula $R - \frac{1}{4}W$.

TABLE 3

*Product-Moment Coefficients between D. A. T. and Otis MA Test
Results and Grades Attained*

	1st Qtr Graphics	2nd Qtr GPA	1st Yr GPA
Mechanical D. A. T.	.31	.16	.22
Spatial D. A. T.	.39	.24	.23
Mechanical + Space	.39	.28	.26
Otis Mental Ability	.35	.36	.35

engineering freshmen at ISU and their first year grade point averages. However, these resulting correlations differ from those presented by Bennett, Seashore, and Wesman (1952) when they con-

sidered the year averages of 90 men in Industrial Arts at State Teachers College, Oswego, New York, and found an r of .10 for Space Relations and .34 for Mechanical Reasoning. They also found correlations of .02 and .09 for space and mechanical respectively when compared with the freshman year averages of 107 Allegheny College males. Bennett et al. report the results of a study made by Berdie, who sampled 57 chemistry and chemical engineering freshman majors at the University of Minnesota Institute of Technology, administered the entire D. A. T. battery, and found, with respect to space and mechanical scores, correlations of .14 and $-.01$ respectively when compared with attained honor point ratios.

It may also be noted that the scores attained on these two tests are relatively well correlated with the grades earned in the graphics course taken during the first quarter.

The question of the reliability of the tests when a 15 minute time limitation is used becomes apparent. The Mechanical Reasoning test lends itself to the use of a split half coefficient of reliability technique. The Spatial test does not afford this type comparison because of the structure of the items. Bennett used the odd-even approach and found a correlation for the Mechanical Reasoning Test to fall between .81 and .88. The correlation coefficient of .82 was found on the results of the Mechanical test when the split-half approach was used on the 15 minute data.

The investigators feel that the 15 minute norms presented in this paper should be quite useful to high school and college counselors when they are dealing with students aspiring to engineering training.

REFERENCE

- Bennett, G. K., Seashore, H. G., and Wesman, A. G. *Manual for the Differential Aptitude Tests*. New York: The Psychological Corporation, 1952.

SOCIAL DESIRABILITY AND THE SEMANTIC DIFFERENTIAL¹

LEROY H. FORD, JR.

AND

MURRAY MEISELS²

State University of New York at Buffalo

THE primary purpose of the present study was to investigate the degree of correspondence between the social desirability variable (Edwards 1957), long viewed as an important factor in personality self-description, and the evaluative dimension (Osgood, Suci, and Tannenbaum, 1957), which consistently emerges as a major factor in semantic differential judgments. The social desirability variable and the evaluative dimension are each receiving considerable attention in the current psychological literature, and each occupies a prominent position in an increasingly large network of empirical data. On the face of it, the two dimensions seem very much alike; each refers to a "good-bad" continuum, and each is regarded as the major source of variation in an important realm of behavior. Yet the two areas of research have developed and continue to flourish quite independently of one another. A few writers (e.g., Feldman and Corah, 1960; Messick, 1960; Zax, Cowen, and Peter, 1963) have indicated an awareness of the similarity between the two dimensions, but their relationship has not been investigated empirically, and no

¹ This paper, in abbreviated form, was read under the title, "Social Desirability and the Evaluative Dimension in Semantic Differential Judgments" at the Eastern Psychological Association convention, Philadelphia, Pa., April 18, 1964.

² The authors wish to express their appreciation to E. P. Hollander, J. W. Julian, and N. N. Markel for their comments on a preliminary draft of the manuscript, and to the staff of the Computing Center of the State University of New York at Buffalo for their assistance in processing the data for this study.

systematic attempt has been made to integrate the data or theoretical formulations associated with them.

The concept of social desirability has been used to refer to: (a) the rated social desirability value of descriptive statements, such as those found in personality and attitude questionnaires; and (b) the differential tendency of individuals or groups to respond to questionnaires or other stimulus situations in a socially desirable manner. The concept of an evaluative dimension has been used to refer to: (a) the evaluative quality or "evaluateness" of the descriptive bipolar scales of the semantic differential, as indexed by the scales' loadings on the evaluative factor; and (b) the evaluative aspect of "meaningful human judgments," i.e., the use of evaluation, by individuals or groups, in the judgment of stimulus objects and events. Thus both concepts have been used in two senses: to refer to (a) a characteristic of items or scales and (b) a characteristic of individual or group behavior. The present investigation is most directly concerned with the relationship between the social desirability value and evaluateness of descriptive items or scales, but the study also relates to the variables of social desirability and evaluation as characteristics of human behavior.

That the concepts of social desirability value and evaluateness are qualitatively similar is suggested by a brief comparison of their content. In social desirability research, personality and attitude statements with high social desirability values are those judged to reflect socially accepted or approved, i.e., "good" characteristics, while items with low values are those judged to reflect socially disapproved, negatively sanctioned, or "bad" characteristics. In semantic differential research, highly evaluative scales are those with high loadings on the evaluative factor; examples are *good-bad*, *kind-cruel*, and *beautiful-ugly*. The social desirability of goodness, kindness, and beauty, and the social undesirability of their opposites seems apparent. Similarly, scales such as *fast-slow* and *thick-thin*, which are evaluatively neutral, also appear to be relatively neutral with respect to social desirability.

For the purposes of this study, the social desirability value of a given bipolar scale was defined as the discrepancy between the mean social desirability ratings of its separate adjectives. The indices of evaluateness were the evaluative factor loadings and dimension coordinates of the scales (Osgood et al., 1957). It was hypothesized

that the concepts of social desirability and evaluativeness are highly similar and that, therefore, high correlations will obtain among the social desirability values, evaluative factor loadings, and evaluative dimension coordinates of the scales. The relationship of social desirability to two other major "dimensions of meaning," potency and activity (Osgood et al., 1957), was also investigated. Since evaluation is largely independent of potency and activity (Osgood et al., 1957), it was predicted that the social desirability values would also be relatively independent of the potency and activity factor loadings and dimension coordinates.

Method

The 50 bipolar scales used by Osgood and his associates in their first two factor analytic studies of the semantic differential (Osgood et al., 1957, pp. 33-46) were chosen as the sample of scales for this study because indices of their evaluativeness were readily available. Social desirability values were obtained for these scales by two different methods (see below), using two different samples of judges. All judges were undergraduate psychology students at the State University of New York at Buffalo. The first set of social desirability values was obtained from 47 judges, 28 male and 19 female; the second set was obtained from 39 judges, 18 male and 21 female. The two sets of values were then correlated with the evaluative, potency, and activity factor loadings and dimension coordinates of the scales as reported by Osgood and his associates (1957, pp. 37 and 43).

Since there is ample evidence that the factor structure of the semantic differential scales is a function of the concept judged (Osgood, 1962; Osgood et al., 1957), it seemed important to have the social desirability ratings made with respect to a concept which was as representative as possible of the concepts used in obtaining the factor loadings. An inspection of the 20 concepts used by Osgood and his co-workers showed that nine (LADY, FATHER, RUSSIAN, ME, BABY, GOD, PATRIOT, MOTHER, and COP) were person or person-like concepts; no category as large or homogeneous as this was apparent in the remaining eleven concepts (BOULDER, SIN, LAKE, SYMPHONY, FEATHER, FIRE, FRAUD, TORNADO, SWORD, STATUE, AMERICA). Accordingly, the social desirability rating instructions specified the concept "people" as the object

of description, and the judges were asked to rate the desirability or undesirability of the adjectives as "human characteristics."

Social Desirability Ratings: Method I

The judges in the first group were each given a mimeographed booklet in which they were to enter their ratings of the adjectives. The instructions and example shown in Table 1 appeared on the

TABLE 1

First Part of Social Desirability Rating Instructions: Method I

On the following pages are a number of adjectives that might be used in describing people. The adjectives are grouped in pairs, with the two adjectives in each pair referring to more or less opposite human characteristics. For each adjective, you are asked to rate, on a scale ranging from "extremely undesirable" to "extremely desirable," how *desirable* or *praiseworthy* you think that characteristic would be considered *from the point of view of our society*.

Record your decision for each adjective by putting a mark (X) on the line next to that adjective at the point that best indicates its desirability or undesirability.

An example of how one person rated the adjectives "intelligent" and "stupid" is given below:

	UNDESIRABLE	DESIRABLE
	Extreme Strong Moderate Mild Neutral Mild Moderate Strong Extreme	
INTELLIGENT	-----	X-----
STUPID	X-----	-----

Note. The complete instructions may be obtained from the first author.

first page of the booklet. These were followed by explanations of the example and some repetition of essential points which, to conserve space, are not given here.³ On the remaining pages of the booklet, the 50 pairs of adjectives were presented, in the manner illustrated in Table 1, with single spacing within pairs and double spacing between pairs. The words designating the nine-point social desirability continuum appeared at the top of each page.

It can be seen that in these social desirability judgments, the two adjectives of each bipolar scale were rated as separate items, whereas, in the semantic differential format, the judges respond to the two adjectives as a bipolar unit. It was in order to preserve, as much as possible, the meaning conferred on the separate adjectives by this bipolar context, that the adjectives were grouped in pairs. In order to further emphasize this point, the instructions for the social desirability ratings specified that the adjectives were

³ A copy of the complete booklet may be obtained from the first author.

"grouped in pairs" and that the two members of each pair referred to "more or less opposite" characteristics.

The judges' ratings were scored on a nine-point scale, ranging from one for "extremely undesirable" to nine for "extremely desirable." For each adjective pair (bipolar scale), the mean rating for each member of the pair, and the difference between the two mean ratings, was determined. This latter value, the difference between the mean ratings of the separate adjectives of a given bipolar scale, was taken as the social desirability value of the scale; and the adjective with the higher (more desirable) mean rating was treated as the positive pole of the scale.

Social Desirability Ratings: Method II

Another difference between presenting the paired adjectives as separate items and presenting them as opposite poles of a single scale is in the degree to which the judge can treat the two adjectives of a given bipolar scale as independent of one another if he wishes to do so. In the social desirability rating method described in the preceding section (Method I), a judge's rating of one member of an adjective pair is free to vary independently of his rating of the other member of the pair. In the semantic differential method, however, the judge is forced to treat the two adjectives of a given scale as polar opposites. It seemed possible that this difference between the social desirability and semantic differential rating formats might introduce a source of "method" variance which could function to lower the obtained correlation between the social desirability values and the indices of evaluativeness.

As a check against this possibility, a second rating method was devised which, following the semantic differential format more closely, placed the two adjectives of each pair at opposite ends of a single graphic scale. The nine scale points ranged from "extremely desirable" at the left, through "neither is more desirable" in the middle, to "extremely undesirable" at the right. Thus the subject could indicate that neither of the adjectives was the more desirable, or he could assign one of four degrees of desirability (mild, moderate, strong, or extreme) to one adjective or the other. The instructions and procedure were otherwise the same as those of Method I except for minor modifications dictated by the bipolar format.⁴

⁴ A copy of the complete booklet used in Method II may be obtained from the first author.

A new sample of 18 male and 21 female judges rated the 50 bipolar scales using the new method. The ratings were scored on a five-point scale ranging from zero for "neither is more desirable" to four for "extremely desirable." For each pair of adjectives, the sum, across judges, of the ratings assigned to each separate adjective was computed, thus taking into account the number of judges who rated that adjective more desirable than the other member of the pair and also the degree of its rated desirability. The difference between the two sums was then divided by the number of judges to yield a mean social desirability value for the pair of adjectives as a bipolar unit. For each adjective pair, the desirable or evaluatively positive pole was taken to be the adjective with the greater summed rating.

Results

The product-moment correlations among the two sets of social desirability values (Method I and Method II) and the evaluative, potency, and activity factor loadings and dimension coordinates are presented in Table 2. The results are given for the combined sexes

TABLE 2
Correlations among Social Desirability Values (SDV) and Evaluative, Potency, and Activity Factor Loadings and Dimension Coordinates

	SDV Method II	Eval. load.	Eval. coord.	Pot. load.	Pot. coord.	Act. load.	Act. coord.
SDV Method I	.99	.92	.88	.15	-.02	.16	.03
SDV Method II		.92	.88	.14	-.04	.17	.06
Eval. load.			.97	.14	-.06	.13	.02
Eval. coord.				.22	.02	.14	.04

Note: $N = 80$, $p < .05$ when $r = .28$, $p < .01$ when $r = .36$.

since the correlations for males and females separately were virtually identical. It will be noted that the social desirability values obtained by the two different methods correlate almost perfectly with one another and show almost identical patterns of correlation with the other variables. It appears, therefore, that the differences in the methods made no appreciable difference in the results. The second method does, however, represent a replication on an independent sample of judges and thus provides evidence for the reliability of the results obtained with the first method. It can be seen

from Table 2 that each of the two sets of social desirability values correlates .92 with the evaluative factor loadings and .88 with the evaluative dimension coordinates, and that all four of these indices are very nearly equivalent in their correlations with the potency and activity factor loadings and dimension coordinates. Thus the results lend strong support to the hypotheses under investigation.

An analysis was also made of the social desirability values obtained by Method I for the separate poles of the bipolar scales. The correlation between these two sets of values was $-.97$ for the combined group of males and females. As might be expected from this finding, the relationship of each of these sets of single-pole values to the evaluative, potency, and activity factor loadings and dimension coordinates was virtually identical to the results obtained with the bipolar values of both Method I and Method II as presented in Table 2.

Discussion

The results of the present study indicate that a large portion of the variance in the evaluative factor loadings and dimension coordinates of the semantic differential scales is predictable from their social desirability scale values, and that the evaluative and social desirability dimensions are both independent of the potency and activity dimensions. Thus the concepts of social desirability value and evaluativeness, as applied to descriptive statements, are highly comparable, if not identical. Although the finding of a substantial degree of correspondence was expected, the magnitude of the relationship was somewhat surprising, at least to the authors, in view of the fact that the evaluative factor loadings represent the result of a complex statistical analysis based on the intercorrelations among judgments on the 50 adjectival scales across 20 diverse concepts, whereas the social desirability values were obtained by a simple and direct rating of the paired adjectives.

Since the several factor loadings and dimension coordinates were based on data obtained from Illinois undergraduates (Osgood et al., 1957), whereas the social desirability values were obtained about eight years later using two independent samples of Buffalo undergraduates, it seems likely that these findings have some generality, at least among college students.

Implications for Theory and Research

In view of the current popularity of both social desirability and semantic differential research, and since even a casual comparison of the two areas would suggest some degree of overlap, it is surprising that the possibility of such overlap has thus far received little attention. Perhaps this is because investigators in the two fields have for the most part addressed themselves to rather different problems. Nevertheless, it should be profitable to examine the research and theory in each of these areas from the vantage point of the other; and the possibility arises that two now relatively discrete but increasingly extensive bodies of knowledge may lend themselves to fruitful integration.

In considering the implications of semantic differential research and theory for the social desirability area, it is pertinent to ask whether the usual interpretation of evaluation as a dimension of "meaning" or "meaningful human judgments" (Osgood et al., 1957) is also applicable to the social desirability variable in personality self-descriptions. A comparison of the instruments most commonly used in assessing these two dimensions, the semantic differential in the case of evaluation, and the personality questionnaire in the case of social desirability, suggests that both instruments tap quite similar behavioral processes. Semantic differential scales and personality questionnaire items are both meaningful verbal stimuli, presented in written form, and used to elicit a set of descriptive or judgmental responses. Just as the semantic differential rater's task is one of making judgments of an explicitly designated concept against a series of descriptive scales, the personality questionnaire respondent's task may be viewed as one of judging an implicit concept such as ME against a series of descriptive statements. Thus, although the two methods differ in format, the responses to both may be regarded as reflecting similar judgmental processes. Therefore, if the semantic differential measures "meaning," so too does the questionnaire; and if evaluation is a dimension of "meaningful human judgments," so too is social desirability.

From this point of view, in spite of the superficial differences between semantic differential scales and questionnaire items, there seems to be no incompatibility in ordering them both to a common evaluative dimension. If, as the combined evidence from social de-

desirability and semantic differential research suggests, evaluation (or, perhaps, "social evaluation") is the major dimension characterizing the realm of descriptive or judgmental statements in general, then the social desirability variable in *self*-descriptive statements may be treated as a special case of this more general dimension. If one wishes to follow Osgood's definition of evaluation as the attitudinal dimension of judgment (Osgood et al., 1957), the social desirability of a questionnaire item or semantic differential scale may be viewed as an index of the cultural "attitude" toward the characteristic referred to by the item or scale; and an individual's endorsement or rejection of a particular characteristic may be taken as one aspect of his "self-attitude."

One implication of the foregoing considerations is that responses to questionnaires and other personality assessment devices may be assumed to tap the same kind of representational mediation process that is hypothesized (Osgood et al., 1957) to underlie semantic differential judgments. A second, related, implication is that the representational model proposed as a link between semantic differential measurement and learning theory (Osgood et al., 1957) may also provide a basis for coordinating personality measurement with learning theory.

Turning attention now to some implications of the social desirability area for work with the semantic differential, it seems likely that many of the results of social desirability research should be applicable to the evaluative variable in semantic differential research. Much of the research in the social desirability area has been concerned with the extent to which individuals describe themselves in accord with perceived social desirability stereotypes rather than giving a "true" description of their feelings and behaviors. It is now well established that personality questionnaires are sensitive to social desirability response "bias," i.e., the tendency to describe oneself in a socially desirable manner (Edwards, 1957). There is also evidence that the magnitude of this effect increases as the social desirability values of the items increase (Edwards and Walsh, 1963). Mention may also be made of the frequently replicated finding (Edwards, 1957) that the group frequency of endorsement of personality questionnaire items is highly correlated with the social desirability values of the items.

The self-descriptive behavior which has provided the data for the

studies referred to above has its most direct semantic differential parallel in subjects' judgments of "self" concepts such as ME, MY ACTUAL SELF, and MY IDEAL SELF. And it is in such semantic differential ratings that the various social desirability phenomena would, presumably, be most likely to appear. Thus, one would expect to find that the tendency to evaluate "self" concepts positively is, in part, a function of the social desirability response tendency as measured by one of the usual social desirability scales (e.g., Crowne and Marlowe, 1960; Edwards, 1957). One would also expect the strength of this relationship to be a function of the social desirability values and evaluative loadings of the bipolar scales against which the judgments are made. And, finally, one would expect both the group frequencies of positive self-evaluation on semantic differential scales, and the mean evaluative scores on the scales, to be highly correlated with the social desirability values and evaluative loadings of the scales.

Although the main body of social desirability research has dealt with personality measurement techniques, there is evidence that the various findings are not confined to trait-descriptive statements or to self-descriptive behavior but extend to attitudinal statements and descriptions of others (Edwards, 1959; Taylor, 1961). Thus there is evidence to suggest that the above hypothesized relationships of social desirability scale values and response bias to semantic differential judgments of "self" concepts may hold for judgments of "other" concepts as well, and the possibility arises that they may also extend to judgments of "non-person" concepts such as, for example, BOULDER, SYMPHONY, or DAWN. In this connection, at least two possibilities seem worth investigating. One is that a highly generalized social desirability tendency, differing among individuals, may be evident across all objects, or a broad class of objects, including the self, other persons, and non-person objects. A second possibility is that an equally general "Pollyanna bias" may occur, i.e., a tendency, differing among individuals, to describe objects favorably or unfavorably, irrespective of the *socially* desirable direction of description. This latter tendency would be confounded with the social desirability tendency when only favorable objects (including the self and desirable others) are judged; but its occurrence can be determined if socially undesirable objects of judgment are also presented. For this reason, the semantic differential lends

itself to the investigation of a personality characteristic which may be confounded in most of the studies to date on social desirability as a response style or personality variable.

Summary

The rated social desirability values of 50 semantic differential bipolar scales were found to be highly correlated with the evaluative factor loadings and dimension coordinates of the scales, and to be largely independent of their activity and potency factor loadings and dimension coordinates. The results indicate that the concepts of social desirability value, as applied to personality questionnaire items, and evaluativeness, as applied to semantic differential scales, are highly comparable, if not identical. Although the possibility has not yet received widespread attention, the research and theory relating to each of these concepts should have many implications for the other. A few of these implications were discussed.

REFERENCES

- Crowne, D. P. and Marlowe, D. "A New Scale of Social Desirability Independent of Psychopathology." *Journal of Consulting Psychology*, XXIV (1960), 349-354.
- Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden, 1957.
- Edwards, A. L. "Social Desirability and the Description of Others." *Journal of Abnormal and Social Psychology*, LIX (1959), 434-436.
- Edwards, A. L. and Walsh, J. A. "The Relationship between the Intensity of the Social Desirability Keying of a Scale and the Correlation of the Scale with Edwards' SD Scale and the First Factor Loading of the Scale." *Journal of Clinical Psychology*, XIX (1963), 200-203.
- Feldman, M. J. and Corah, N. L. "Social Desirability and the Forced Choice Method." *Journal of Consulting Psychology*, XXIV (1960), 480-482.
- Messick, S. "Dimensions of Social Desirability." *Journal of Consulting Psychology*, XXIV (1960), 279-287.
- Osgood, C. E. "Studies on the Generality of Affective Meaning Systems." *American Psychologist*, XVII (1962), 10-28.
- Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning*. Urbana: University of Illinois Press, 1957.
- Taylor, J. B. "What Do Attitude Scales Measure: The Problem of Social Desirability." *Journal of Abnormal and Social Psychology*, LXII (1961), 386-390.
- Zax, M., Cowen, E. L., and Peter, Sister Mary. "A Comparative Study of Novice Nuns and College Females Using the Response Set Approach." *Journal of Abnormal and Social Psychology*, LXVI (1963), 369-375.

VALIDITY STUDIES SECTION

WILLIAM B. MICHAEL, Editor

University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor

Centinela Valley Union High School District

and the

University of Southern California

<i>Validity of the Durrell-Sullivan Reading Capacity Test.</i> HARRY SINGER	479
<i>WAIS and Group Test Predictions of an Academic Success Criterion: High School and College.</i> ROBERT CONRY AND WALTER T. PLANT	493
<i>The Interaction of Ability Levels and Socioeconomic Variables in the Prediction of College Dropouts and Grade Achieve- ment.</i> BEN BARGER AND EVERETTE HALL	501
<i>Need Patterns and Abilities of College Dropouts.</i> JAY L. CHAMBERS, BEN BARGER, AND LEWIS R. LIEBERMAN	509
<i>Two Cross Validations of the Opinion, Attitude and Interest Survey.</i> SAM C. WEBB	517
<i>Validation of a Carefulness Test Battery.</i> H. G. OSBURN, CECIL J. MULLINS, AND DANIEL E. SHEER	525
<i>An Evaluation of an Attitude Scale toward Teaching.</i> PAUL L. CRAWFORD	535
<i>Measurement of Achievement Motivation in Army Security Agency Foreign Language Candidates.</i> WILLIAM E. DATEL, FORREST D. HALL, AND CHARLES P. RUFÉ	539
<i>American College Test (ACT) Performance as a Function of Examinee Acceptance of Test.</i> HENRY F. DIZNEY, ELINOR A. ELFNER, AND HORACE A. PAGE	547

- Predicting Grade Point Average with the SRA Tests of Educational Ability: A 13-Month Follow-Up Study.* WARREN S. BLUMENFELD 555
- The Prediction of Grades in Introductory Psychology from Tests of Primary Mental Abilities.* BISHWA NATH MUKHERJEE 557
- Intellective Predictors of Success in Nursing School.* JON M. PLAPP, GEORGE PSATHAS, AND DANIEL V. CAPUTO 565
- The Predictive Validity of a Battery of Diversified Measures Relative to Success in Student Nursing.* WILLIAM B. MICHAEL, RUSSELL HANEY, AND STEPHEN W. BROWN 579
- Correlates of Achievement on the Admissions Test for Graduate Study in Business.* ARTHUR MITTMAN AND JOHN W. LEWIS 585
- The Peabody Picture Vocabulary Test in Comparison with Other Intelligence Tests and an Achievement Test in a Group of Mentally Retarded Boys.* FRANCES M. THRONE, JOSEPH C. KASPAR, AND JEROME L. SCHULMAN 589
- PMA Factors, Sex, and Teacher Nomination in Screening Kindergarten Gifted.* PHILLIP WEISE, C. E. MEYERS, AND JOHN K. TUEL 597
- The Application of a Configuration Method to the Prediction of Success in First Grade.* JEAN L. BALINKY 605
- The Predictive Value of a Beginning First-Grade Intelligence Examination.* CARRIE M. SCOTT 613

VALIDITY OF THE DURRELL-SULLIVAN READING CAPACITY TEST¹

HARRY SINGER

University of California, Riverside

THE *Durrell-Sullivan Reading Capacity Test (DSRC)*, designed to measure reading capacity in grades 3 to 6, is based on the principle that the potential reading achievement of an individual should be equal to his auditory comprehension.² The principle is dependent upon a basic assumption of uniformity in brain functioning in response to language relationships, whether input is through the visual or through the auditory system. This basic principle and its neurological assumption were explicitly formulated by Sullivan (1938, p. 40) in the first description of the test:

The principle underlying the use of measures of auditory comprehension as criteria for potential reading achievement is that if the mind is able to handle auditory symbols up to a certain degree of complexity, it should be able to handle visual symbols up to that same degree of difficulty. This principle, of course, assumes a uniformity of brain structure in regard to the handling of symbolic relationships that are involved in language.

Although *DSRC* has now been in use for some 25 years, there is

¹ This investigation was part of a larger study supported by the U. S. Office of Education, Department of Health, Education and Welfare, Contract No. 2011.

² *DSRC* has two subtests. (1) For *Word Meaning*, the tester pronounces a word which corresponds to one of a group of eight pictures. (2) On *Paragraph Meaning*, the tester reads a short story, then asks five questions about the story. The pupil answers each question by selecting an appropriate picture from a set of three pictures. The reliabilities for these subtests are high: for grades 3 to 6 they range from .90 to .96 for *Word Meaning* and from .83 to .93 for *Paragraph Meaning* (Sullivan, 1938).

still very little evidence to support its validity as a measure of reading potential.

Related Research. Whatever *DSRC* measures does improve with grade level because there is an average increment of 16 points of raw scores between each year level (Alden, Sullivan, and Durrell, 1941). A correlation between *DSRC* and *Stanford-Binet Intelligence Test (SB)* scores for 80 children in grades 4 through 7 enrolled in a public school remedial reading program was .76, but *DSRC* in comparison with *SB* overestimated reading potential (Bliesmer, 1956). In a multiple regression equation, *DSRC* had about equal weight with the *California Test of Mental Maturity* in predicting scores on *California Reading Test* (Owen, 1958). However, for a sample of 87 fourth graders the correlation between *DSRC* and *Gates Basic Reading Tests Types A and D*, ranged from only .41 to .54; the highest correlation between *DSRC* and *Primary Mental Abilities (PMA)* was .64 between *DSRC* Word Meaning and *PMA* Pictures (Bond and Clymer, 1955). Toussaint (1961) found that the *STEP* Listening Test had a closer relationship with *Gates Reading Survey* than did *DSRC*, but the *STEP* test has some reading in it and therefore its correlation with the *Gates Reading Survey* may be spuriously high. The *DSRC*, although of some use in the clinic, is highly limited in estimating potential reading capacity of children with foreign language backgrounds or other language handicaps (Robinson, 1953).

Problem. The general purpose of this study is to test by means of a factor-analysis model the basic assumption underlying *DSRC* that there is uniformity of brain functioning in response to language relationships. Therefore, the following specific questions were formulated. (a) What is the factor analytic structure of *DSRC* when it is embedded in a matrix of variables selected for their known ability to predict speed and power of reading (Singer, 1960, 1962)? (b) Is the factor-loading pattern of *DSRC* similar to that of speed and power of reading? (c) How similar are the factor loading patterns of *DSRC* and such subskills as word meaning, word recognition, and visual and auditory perceptual abilities?

Further clarification of these questions is necessary. If there is a "uniformity of brain structure" in handling the symbolic relationships involved in responding appropriately to visual symbols of reading tests and to auditory symbols of *DSRC*, then the factor ana-

lytic structure of the visual and auditory tests would be similar. That is, the same factors or mental functions would contribute to the variability of the tests, if not equally then at least proportionately. If so, *DSRC* would gain support as a valid measure of potential reading achievement.

However, reading ability is not unidimensional, but divides into two major interrelated components, speed and power of reading. Underlying and supporting each component is a complexly interrelated structure of subskills and capacities (Holmes, 1948). Broadly categorized, this general structure consists of interrelated input, mediating, output, and both short-term and long-term memory systems; all of these systems are overlaid with emotional systems and undergirded by physiological systems (Holmes and Singer, 1964; Davis, 1964). Within the limitations of developmental changes and test battery comparability, this general structure for attaining speed and power of reading has been verified at the college (Holmes, 1954a), high school (Holmes and Singer, 1961), and intermediate grade level (Singer, 1962).

These subskills and capacities are predictors at some level in the general structure for speed and power of reading. For example, the following are some of the predictors which occur in the structural model for power of reading in the fourth grade (Singer, 1960, 1962). At the lowest level, spelling recognition together with prefixes and spelling recall enter into the constellation of subabilities that make up Word Recognition in Context. At the middle level, word recognition in context, plus suffixes, and mental age contribute to the variance in Vocabulary in Isolation. Finally, on the highest level, vocabulary in isolation becomes integrated with suffixes, mental age, and matching sounds in words to culminate in Power of Reading.

The question then is whether *DSRC* is also a "capacity" test for one or more of these predictors, particularly the word recognition predictors. This question is quite important because at least from an instructional viewpoint the nature of the reading task changes during the developmental continuum. In the initial stages of reading instruction, development of perceptual and word recognition subskills is emphasized. At this stage, individuals have already matured sufficiently in their reasoning or mediational processing systems so that they could adequately comprehend the relatively simple ideas

presented in beginning instructional material, provided that their input or word recognition subsystem were adequately developed for transforming printed stimuli into mental processes. But, during this initial stage there are individual differences in the input system which may be attributable to variation in ability to conceptualize linguistic stimuli (Singer, 1960), effectiveness of instructional strategies, modality sensitivity and receptivity, or to an interaction of these sources of variance (Bond, 1935; Fendrick, 1935). However, as individuals progress through the grades, they gradually tend to master word recognition processes (Singer, 1964); instructional emphasis then shifts to further development of ability to reason about the increasingly complex ideas presented in the instructional material. Hence, during the developmental continuum of learning to read, there is a shift in instructional emphasis from an estimate of input to an estimate of mediational processing potential.

Method

Sample. A battery of 30 tests was administered to 283 fourth graders in a school located in an average socio-economic district in Alvord, California. From comparison of the means of the sample data with standardized test norms on age, I.Q., Speed and Power of Reading, as shown in Table 1, the sample appears to be somewhat representative of the general population of fourth graders because it is close to the norms on I.Q.; *Gates Reading Survey*, Speed of Reading; and *Gates Reading Survey*, Level of Comprehension or Power of Reading.

The grade equivalents for the current sample, according to the norms in the Durrell-Sullivan Manual, is 5.8 on Paragraph Meaning *Achievement* and 5.4 on Paragraph Meaning *Capacity*. Not only is the sample higher on achievement than on capacity, but the sample is also advanced approximately one grade level on both tests! However, on the *Gates Reading Survey* results, the current sample is approximately at grade level. A similar comparison between *DSRC* and *Gates Reading Survey* yielded comparable results in a previous investigation (Singer, 1960). These findings suggest that the Durrell-Sullivan norms probably overestimate grade equivalencies.

The cumulative records of the subjects revealed that they had been taught by a wide variety of teachers, had used a heterogeneous set of basal and supplementary readers, and had been regis-

TABLE 1

Comparison of Certain Sample Means with Fourth Grade Norms

Variable	Sample Mean N = 283	National Norm
Chronological Age	9-11	9-11
PMA Intelligence Quotient	102.0	100.0
Gates Speed of Reading	19.8	18.7
Gates Level of Comprehension	15.6	18.0

tered in many school systems throughout the country. Therefore, the results of this study cannot be related to any particular set of materials nor to any particular methodological emphasis.

Tests. A test battery, listed in Table 2, was constructed of variables which would presumably measure comparable input and mediational processes in the visual and auditory systems for reading and listening, respectively. The tests were selected from the batteries of Durrell and Sullivan (1937); Gates (1953); Holmes (1962); Kwalwasser-Dykema-Holmes Test, Holmes (1954b); Singer (1963); Thurstone and Thurstone (1954); and Van Wagenen and Dvorak (1953).

Reliability coefficients, also presented in Table 2, reveal that all the tests had substantially high reliabilities. Bivariate distributions of each variable with Speed of Reading and Level of Comprehension, respectively, satisfactorily passed the chi-square test for rectilinear regression.³

Concurrent validity coefficients between each of the tests and the subtests of *DSRC* are also given in Table 2. The highest correlation is .64 between *DSRC* Word Meaning and *PMA* Pictures, exactly the same degree of relationship between these variables as that reported by Bond and Clymer (1955). The next highest correlation is .56 between the subtests of *DSRC*, which means that listening vocabulary and listening comprehension in this sample have only 31 percent variance in common. The correlation of .48 between *Durrell-Sullivan* Paragraph Meaning Capacity and Paragraph Meaning Achievement is surprisingly low, since Sullivan (1938) stated that these tests were constructed with parallel content and

³ Computations were performed on the IBM 7090 Computer at the University of California, Berkeley, by means of RSCAT program written by M. Maruyama. Mr. Price Stiffler, programmer consultant, is acknowledged for his extremely competent assistance in processing all the data for this study on the computer.

TABLE 2
Statistical Data on 50 Variables for 283 Fourth Graders

Test Battery	r ₁₁	Correl. with Capacity for: Words	VVM	Principal Components Rotated Factor Loadings Factors*				
				1	2	3	4	5
				VVM	AUD	VR	SVP	AP
1 Durrell-Sullivan Reading Capacity								
Word Meaning	85**	—	56	18**	75	18	20	16
Paragraph Meaning	87	56	—	16	62	40	10	25
2 Gates Reading Survey								
Speed of Reading	88	40	33	62	39	-19	35	15
Level of Comprehension	89	45	45	75	40	12	-02	13
3 Durrell-Sullivan Reading Achievement								
Paragraph Meaning	91	52	48	66	44	11	16	24
4 Thurstone Primary Mental Abilities								
Words	91	43	38	76	39	00	00	15
Pictures	70	64	49	28	78	10	07	11
Space	83	25	42	14	19	74	-01	12
Word Grouping	79	39	42	68	29	30	02	08
Figure Grouping	84	37	42	20	21	73	09	06
Perception	94	28	34	27	04	57	40	07
5 Van Wagenen-Dvorak Silent Reading								
Range of Information	73	49	53	40	67	13	04	04
6 Singer Linguistic Tests								
Auditing Conceptual Ability	76	45	49	51	34	29	10	25
Meaning of Affixes	77	46	46	71	43	12	04	22
Word Recog. in Context	93	28	34	80	11	18	-02	15
Matching Sounds in Words	96	35	32	86	18	17	-01	03
Blending Word Elements	85	32	32	80	11	23	05	12

TABLE 2—Continued
Statistical Data on 30 Variables for 283 Fourth Graders

Test Battery	r ₁₁	Correl. with Capacity for: Words Par.	Principal Components Rotated Factor Loadings Factors*				
			1 VVM	2 AUD	3 VR	4 SVP	5 AP
Phonics	92	38	75	24	16	02	16
Syllabication Consistency	83	28	71	11	15	13	11
Auditory Verbal Abstraction	90	30	75	14	19	00	09
Spelling Recognition	90	34	83	16	00	19	13
Speed of Word Perception	80	29	57	11	-00	46	14
Recog. of Affixes and Roots	89	31	68	02	24	14	10
Word Reversals	73	30	43	06	48	15	36
7 Holmes Language Perception							
Word Embedded	92	24	59	03	17	44	-01
Figure and Ground	78	20	-01	08	-01	82	-04
Cue Symbol Closure	78	29	07	16	29	65	10
8 K-D-H Musical Aptitudes							
Pitch Discrimination	73	21	17	21	10	-01	50
Rhythm Discrimination	64	24	19	01	15	07	72
Tonal Intensity Discrim.	82	24	07	15	02	02	73

*Identification of Factors

1. Visual Verbal Meaning
2. Auding
3. Visual Relationships
4. Speed of Visual Perception
5. Auditory Perception

**Decimals before correlations and factor loadings have been omitted.

comprehension questions. At the correlational level then, *DSRC* subtests are not highly predictive of any of the variables used in this study.

Factor Analysis. A principal components factor analysis with communalities of 1.0 was used to factor the matrix. The rank of the matrix was specified as the number of eigenvalues equal to or greater than 1.0. Kaiser's (1959) normalized varimax rotation technique for maximum interpretability was employed.⁴

Results and Interpretation

The rotated principal component factor loadings, shown in Table 2, yielded five interpretable factors. Factor I was identified as *Visual Verbal Meaning* because tests with high loadings on this factor require subjects to read for comprehension, vocabulary, and word recognition. Factor II was labeled *Auding* since the listening tests, such as *PMA Pictures*, *DSRC* subtests, and Range of Information correlate highly with this factor. Factor III was named *Visual Relationships* to represent its saturation of *PMA Space*, Figure Grouping, and Perception, plus its substantial correlations with *DSRC* subtests and Word Reversals. Factor IV was defined as *Speed of Visual Perception* by high test loadings of Speed on Reading, Perception, Speed of Word Discrimination, Word Embedded, Figure and Ground, and Cue Symbol Closure. Factor V was called *Auditory Perception* because of its high correlations with Pitch, Rhythm, and Intensity.

Comparison of the factor loadings of either the *Gates* or *Durrell-Sullivan* reading comprehension tests with either of the *DSRC* subtests reveals that their patterns are not similar. The reading comprehension tests correlate .62 to .75 with *Visual Verbal Meaning* and .40 to .44 with *Auding Factors*. The *DSRC* subtests' highest loadings are .62 to .75 on *Auding* and .18 to .40 on *Visual Relationships Factors*. Although both the reading and the listening tests have substantial loadings on Factor II, the quantitative variation in their factor pattern does not substantiate the assumption that brain functioning in performance on these tests is uniform. On the contrary, the evidence supports the contention that the visual and auditory

⁴Dr. Alan B. Wilson, Survey Research Center, University of California, Berkeley, wrote the principal components factor analysis program, FA80, and integrated it with the varimax program.

systems mobilized for performance on the reading and listening tests, although having some common functions and therefore some degree of cortical interfacilitation, are nevertheless separate systems. Hence, at least at the fourth grade level, listening comprehension alone cannot justifiably be used as a valid measure of concurrent reading achievement, and vice versa.

Nor should *DSRC* subtests be used as a valid measure of concurrent word recognition achievement at the fourth grade level because none of the word recognition measures has any substantial loading on the *Auding Factor*. Furthermore, the loadings of .18 for *DSRC* Paragraph Meaning and .16 for *DSRC* Word Meaning on the *Visual Verbal Meaning Factor* are quite low. Again, the evidence suggests that two more or less separate systems are operating in performance on *DSRC* subtests and on word recognition abilities.

Discussion

If the *DSRC* type of test is taken as a valid measure of an individual's reading potential, then an explanation has to be sought for a significant discrepancy between reading potential and reading achievement, even when achievement is actually higher than potential (Alden, Sullivan, and Durrell, 1941; Maxwell, 1959). For some individuals the discrepancy may be validly attributed to inadequate instruction, desire to learn, or to some other causal factor, all of which assume that under optimal conditions there would be no discrepancy. A more general explanation, supported by the results of this study and by the findings of a similar investigation on only 60 fourth graders (Singer, 1960), is that at least two separate, though moderately interrelated systems are mobilized for performance in reading and in listening; therefore, an individual could perform better or have higher potential in one than in the other system, possibly as a result of intraindividual variation in mental capacities or asynchronous development of mental functions (Bayley, 1949).

Further support for the interpretation of the separateness of the two systems can be adduced from several studies: Gates (1926) concluded that visual perception for words, objects, and geometric symbols are specific abilities; Karwoski, Gramlick, and Arnott (1944) inferred that the longer reaction times for objects and pictures than for words was due to the formulation of an intermediary symbol before a verbal response to objects and pictures could be

made; Strang (1943) explained that verbal and nonverbal mental tests tap different mental processes; Caffrey (1953) at the high school level identified an auding factor, which was distinct from reading comprehension, mental age, chronological age, and interests; and Spearitt (1962) at the sixth grade level using the *STEP* listening test in a battery that included reading, reasoning, and rote memory, also isolated a listening comprehension factor. Consistent with all these findings is the localization theory of neurology (Nielsen, 1951) with its implications for the reading process (Holmes, 1957) that different areas of the brain are involved in (a) *visual* perception of objects, pictures, and words and (b) *auditory* perception of music and language. Moreover, from a battery of tests the best predictor of first grade reading achievement was the visual word discrimination subtest of *Gates Reading Readiness* (Balow, 1963). It would therefore seem that a valid test of silent reading potential would necessarily be weighted with items or scales that require perception, retention, manipulation and conceptualization of *written* or *printed* verbal symbols, with input through the visual mode.

Summary

The validity of the basic assumption supporting the use of the *Durrell-Sullivan Reading Capacity Test* as a measure of reading potential was investigated by means of principal components factor analysis. Factors were extracted from a matrix of 30 variables that had been selected to measure both visual and auditory input and mediational processing systems for listening and for reading. The varimax rotated factor loadings for a sample of 283 fourth graders did not support the Durrell-Sullivan assumption that there is a uniformity of brain structure in regard to the handling of symbolic relationships in listening and reading tests. The listening tests primarily loaded on an *Auding Factor* while the reading tests primarily tapped a *Visual Verbal Factor*. However, the pattern of loadings suggested that what *DSRC* actually assesses in the fourth grade is listening comprehension at a concrete or auditory-visual associational level rather than listening comprehension at a more abstract level. Consequently, an alternate hypothesis was advanced that what is mobilized for performance in listening and reading in the fourth grade are two separate, though moderately interrelated, mul-

tidimensional systems in which individuals could have higher potential in one system than in the other. Caution should therefore be exercised in the use in the fourth grade of the *Durrell-Sullivan Reading Capacity Test* alone for assessing concurrent reading capability.

This conclusion should not, of course, be generalized to other measures of listening comprehension, to other grade levels, nor to other curricula without further investigation. In fact, Holmes and Singer (1961) found in a factor analytic study of a similar battery at the high school level that another measure of listening comprehension and reading achievement did indeed correlate highly with the same factor. Further integration in these two systems apparently occurs sometime after the fourth grade level. Moreover, it is possible that emphasis at the elementary level upon the development of listening comprehension may accelerate this integration. If so, then listening comprehension would serve as a more valid group estimate of concurrent reading capability even at the elementary school level, but caution would still be necessary in estimating expectancy levels in particular individuals because of the possibility of (a) intraindividual variation in capacities or (b) asynchronous development of an individual's cognitive systems for listening and for reading.

REFERENCES

- Alden, Clara L., Sullivan, Helen B., and Durrell, D. D. "The Frequency of Special Reading Disabilities." *Education*, LXII (1941), 32-36.
- Balow, I. H. "Sex Differences in First Grade Reading." *Elementary English*, XL (1963), 303-306; 320.
- Bayley, Nancy. "Consistency and Variability in the Growth of Intelligence from Birth to Eighteen Years." *Journal of Genetic Psychology*, LXXV (1949), 165-196.
- Bliesmer, E. P. "A Comparison of Results of Various Capacity Tests Used with Retarded Readers." *Elementary School Journal*, LVI (1956), 400-402.
- Bond, G. L. "The Auditory and Speech Characteristics of Poor Readers." *Teachers College Contributions to Education*, 1935, No. 657.
- Bond, G. L. and Clymer, T. W. "Interrelationship of the SRA Primary Mental Abilities, Other Mental Characteristics, and Reading Ability." *Journal of Educational Research*, XLIX (1955), 131-136.
- Caffrey, J. P. "Auding Ability as a Function of Certain Psychometric Variables." Unpublished doctoral dissertation, University of California, Berkeley, 1953.

- Davis, F. R. "The Substrata-Factor Theory of Reading: Human Physiology as a Factor in Reading." In J. A. Figurel (Ed.), *Improvement of Reading through Classroom Practice*, Proceedings of the Ninth Annual Convention of the International Reading Association, IX (1964), 292-296.
- Durrell, D. D. and Sullivan, Helen B. *Manual for Durrell-Sullivan Reading Capacity and Reading Achievement Tests*. New York: World Book, 1937.
- Fendrick, P. "Visual Characteristics of Poor Readers." *Teachers College Contributions to Education*, 1935, No. 656.
- Gates, A. I. "A Study of the Role of Visual Perception, Intelligence, and Certain Associative Processes in Reading and Spelling." *Journal of Educational Psychology*, XVII (1926), 433-445.
- Gates, A. I. *The Manual of Directions for Gates Reading Survey*. New York: Teachers College, Columbia University, Bureau of Publications, 1953.
- Holmes, J. A. "Factors Underlying Major Reading Disabilities at the College Level." Unpublished doctoral dissertation, University of California, Berkeley, 1948.
- Holmes, J. A. "Factors Underlying Major Reading Disabilities at the College Level." *Genetic Psychology Monographs*, XLIX (1954), 3-95. (a)
- Holmes, J. A. "Increased Reliabilities, New Keys, and Norms for a Modified Kwalwasser-Dykema Test of Musical Aptitudes." *Journal of Genetic Psychology*, LXXV (1954), 65-73. (b)
- Holmes, J. A. "The Brain and the Reading Process." In *Reading Is Creative Living, Twenty-Second Yearbook of Claremont College Reading Conference*. Claremont, California: Claremont College Curriculum Laboratory, 1957. Pp. 49-67.
- Holmes, J. A. *California Language Perception Tests* (Revised). Palo Alto: Educational Development Corporation, 1962.
- Holmes, J. A. and Singer, H. "The Substrata-Factor Theory: Substrata Factor Differences Underlying Reading Ability in Known-Groups." Final report covering contracts 538 and 538A, Office of Education, U. S. Department of Health, Education, and Welfare, 1961.
- Holmes, J. A. and Singer, H. "Theoretical Models and Trends toward More Basic Research in Reading." *Review of Educational Research*, XXXIV (1964), 127-155.
- Kaiser, H. F. "Computer Program for Varimax Rotation in Factor Analysis." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIX (1959), 413-420.
- Karwoski, T. F., Gramlick, F. W., and Arnott, P. "Psychological Studies in Semantics: I. Free Association Reactions to Words, Drawings, and Objects." *Journal of Social Psychology*, XX (1944), 233-247.
- Maxwell, J. "Durrell-Sullivan Reading Capacity and Achievement." In O. K. Buros, (Ed.), *Fifth Mental Measurements Yearbook*, V (1959), 661-662.

- Nielsen, J. M. *A Textbook of Clinical Neurology*. New York: Paul B. Hoeber, 1951.
- Owen, J. C. "A Study of the Prognostic Value of Certain Measures of Intelligence and Listening Comprehension with a Selected Group of Elementary Pupils." Unpublished doctoral dissertation, University of Missouri, Columbia, Missouri, 1957. (*Dissertation Abstracts*, XIX (1958), 484.)
- Robinson, Helen M. "Durrell-Sullivan Reading Capacity and Achievement Tests." In O. K. Buros, (Ed.), *Fourth Mental Measurements Yearbook*, IV (1953), 562-564.
- Singer, H. "Conceptual Ability in the Substrata-Factor Theory of Reading." Unpublished doctoral dissertation, University of California, Berkeley, 1960.
- Singer, H. "Substrata-Factor Theory of Reading: Theoretical Design for Teaching Reading." In J. A. Figurel, (Ed.), *Challenge and Experiment in Reading*, Proceedings of the Seventh Annual Convention of the International Reading Association, VII (1962), 226-232.
- Singer, H. "California Linguistic Tests, Elementary School Level" (Revised). Riverside: University of California, 1963. (Multilith)
- Singer, H. "Substrata-Factor Theory of Reading: Grade and Sex Differences in Reading at the Elementary School Level." In J. A. Figurel, (Ed.), *Improvement of Reading through Classroom Practice*, Proceedings of the Ninth Annual Convention of the International Reading Association, IX (1964), 313-320.
- Spearitt, D. "Listening Comprehension: A Factorial Analysis." Australian Council for Educational Research, (Melbourne, Victoria), 1962. Summarized in David A. Russell, "A Conspectus of Recent Research on Listening Abilities." *Elementary English*, XLI (1964), 3, 262-267.
- Strange, Ruth. "Relationship between Certain Aspects of Intelligence and Certain Aspects of Reading." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, III (1943), 355-359.
- Sullivan, Helen Blair. "A New Method of Determining Capacity for Reading." *Education*, LIX (1938), 39-45.
- Thurstone, L. L. and Thurstone, Thelma G. *SRA Primary Mental Abilities, For Ages 7 to 11*. Chicago: Science Research Associates, 1954.
- Toussaint, Isabella H. "Interrelationships of Reading, Listening, Arithmetic, and Intelligence and Their Implications." Unpublished doctoral dissertation. Pittsburgh: University of Pittsburgh, 1962. (*Dissertation Abstracts*, XXII (1961), 819.)
- Van Wagenen, M. J. and Dvorak, A. *Diagnostic Examination of Silent Reading Abilities*. Minneapolis: M. J. Wagenen, 1953.

WAIS AND GROUP TEST PREDICTIONS OF AN ACADEMIC SUCCESS CRITERION: HIGH SCHOOL AND COLLEGE¹

ROBERT CONRY² AND WALTER T. PLANT

San Jose State College

The purposes of this study were: (a) to determine the predictive validity of all parts of the *Wechsler Adult Intelligence Scale: WAIS* (Wechsler, 1955) for educationally relevant criteria in high school and in college, and (b) to determine whether the best predictor subparts of the *WAIS* correlated with the criterion measures as high as did scores from widely used group tests of academic aptitude.

It has become common to find the *WAIS* used in secondary school counseling and guidance centers, in community vocational rehabilitation centers, and in college counseling centers and remedial study programs. Invariably, predictions of future educational success are made on the basis of *WAIS* results obtained in these settings. Yet, little in the way of predictive validity information for educational success criteria is available to those using the *WAIS*.

Three years after the publication of the *WAIS*, Wechsler (1958) published information on the relationship between *WAIS* scores and educational attainment for the *WAIS* standardization sample, but there have been only two predictive validity studies relevant to our purposes here. Plant and Lynd (1959) correlated the *WAIS* Verbal, Performance, and Full-Scale weighted scores with college

¹ Adapted from a thesis completed by the first author under the direction of the second author and submitted to the Department of Psychology, San Jose State College, in partial fulfillment of the requirements for the bachelor of arts (Honors) degree.

² Now an NDEA Fellow, Department of Educational Psychology, University of Wisconsin.

freshman grade point average for a sample of 161 subjects tested with the *WAIS* as college freshmen. They reported correlations of from .31 to .58 for the weighted *WAIS* scores and the criterion. Plant and Lynd also correlated Linguistic, Quantitative, and Total raw scores of the *American Council on Education (ACE) Psychological Examination* (Thurstone and Thurstone, 1948) and the freshman year grade point average criterion for their 161 subjects, and obtained coefficients ranging from .18 to .46.

A more recent study (Wall, Marks, Ford, and Ziegler, 1962) was completed with a random stratified sample of 106 Pennsylvania State University entering freshmen. Wall *et al.* correlated *WAIS* sub-scale scores and scores from group aptitude tests with first semester grade point average for males and females and for science and non-science majors separately. They reported validity coefficients of about .30 for the *WAIS* Verbal Scale and Full-Scale scores and a range in coefficients of .33 to .42 for the group measures.

In each of the two studies, it was concluded that criteria of academic success in college were as well predicted from *WAIS* scores as from group academic aptitude test scores. In neither study were correlations reported for the 11 subtests of the *WAIS*; nor have the writers found predictive validity studies involving educationally relevant criteria and the *WAIS* for less select and less homogeneous samples than samples of college students. The presentation of results of correlational analyses from two samples will provide validity data for all parts of the *WAIS*, and for one sample that is less select than is a college student sample.

Procedure

Study I: High School

Subjects. The high school sample was comprised of 98 students, each of whom was a junior at James Lick High School in San Jose, California at the time of testing with the predictor tests. Subjects were selected by a stratified random sampling procedure. All juniors at the study school were assigned originally by school authorities to one of three types of English classes. One of each of the three types of English classes was selected at random, and all students in each of these three classes were tested with the *WAIS*.

In the sample of 98 there were 60 females and 38 males, but in

the interest of maintaining adequate sample size, all subjects were combined into a single group for the analyses to be reported.

Predictors. The *WAIS* was administered to subjects by college seniors and first-year graduate students enrolled in a practicum course in intelligence testing under the direction of the second author. The *WAIS* tests were administered according to the directions in the test manual, and all 11 subtests were administered to each subject. Each test protocol was checked for scoring accuracy by a teaching assistant, and all testing was accomplished while the subjects were in the early quarter of the eleventh grade.

At the beginning of the junior year of the study high school, the *Terman-McNemar Test of Mental Ability* (Terman and McNemar, 1941) and the *Chicago Test of Primary Mental Abilities Test* (Thurstone and Thurstone, 1953) were administered to all enrolled juniors. Composite scores for these tests were obtained from school officials for each of the 98 subjects in the study sample. Neither test was used as a basis for sectioning the subjects into the English classes.

Criterion. The criterion selected for high school academic success was the rank in the graduating class. The ranks for all members of the graduating class were converted to C-scale standard scores (Guilford, 1956), and the C-scale standard score for each of the 98 study subjects constituted the criterion score. In effect, this score was approximately a 20-month post-test criterion measure.

Study II: College

Subjects. The college sample was comprised of 335 subjects, all freshmen at San Jose College at the time of testing with the predictor tests. Subjects were volunteers obtained from introductory psychology classes over several years in a college in which a course in general psychology is required for completion of the bachelor's degree.

In this sample of 335 there were 188 females and 147 males. Separate analyses were made for sex groups (Conry, 1963), but in the interest of being consistent with the report for the high school sample, all subjects were combined into a single group for the analyses to be reported here.

Predictors. The *WAIS* was administered to subjects by college seniors and first-year graduate students enrolled in a practicum

course in intelligence testing under the direction of the second author. The *WAIS* tests were administered according to the directions in the test manual, and all 11 subtests were administered to each subject. Each test protocol was checked for scoring accuracy by a teaching assistant, and all testing was accomplished while the subjects were in their freshman year at the college.

At the time of obtaining *WAIS* data for this study, the *ACE Psychological Examination* (Thurstone and Thurstone, 1948) test was routinely administered to all entering freshmen at the college. Scores for this test were obtained from the College Testing Office for each of the 335 subjects in the study sample, and were converted to C-scale standard scores because the scores were obtained from three different forms of the *ACE* test.

Criterion. The criterion selected for college academic success was the grade point average, on a four point scale, at the end of the freshman year. No attempt was made to divide the sample into groups with different academic majors as was done in the study by Wall *et al.* (1962), nor was any attempt made to predict course grades in any particular subject matter. The criterion measure was the total freshman grade point average (GPA). In effect, this measure was approximately a six-month post-test criterion score.

Analyses: Both Samples

Predictive validity coefficients for both samples were obtained by conventional Pearsonian correlational methods. For Study I, 19 test scores (11 *WAIS* subtest scaled scores, *WAIS* Verbal, Performance, and Full-Scale summed scaled scores and I.Q.'s, and the composite scores from each of the two group tests) were correlated with the rank-in-graduating class C-score and with each other. For Study II, 20 test scores (17 *WAIS* scores as above, and the Quantitative, Linguistic, and Total *ACE Psychological Examination* scores converted to C-scale scores) were correlated with the freshman GPA criterion and with each other.

Statistical tests of the significance of the difference between correlated correlation coefficients (Walker and Lev, 1953) were computed for both studies for: (a) the highest *WAIS* subtest and criterion coefficient vs. the highest group test and criterion coefficient, and (b) the best *WAIS* composite score and criterion coefficient vs. the best group test and criterion coefficient.

Results

The means and standard deviations for the *WAIS* Verbal, Performance, and Full-Scale I.Q.'s, the group tests, and the criterion scores for both study samples are found in Table 1.

TABLE 1

*High School and College Means and Standard Deviations for
WAIS I.Q.'s, Group Tests, and Criterion Scores*

Variate	H.S. <i>N</i> = 98		Coll. <i>N</i> = 335	
	Mean	σ	Mean	σ
WAIS: Verbal I.Q.	106.1	14.2	115.1	8.4
Performance I.Q.	107.3	11.2	112.6	9.6
Full-Scale I.Q.	107.1	12.4	114.8	8.0
Terman-McNemar I.Q.	103.2	17.9		
Primary Mental Abilities	100.3	15.3		
Graduating Rank C-score	6.2	2.3		
ACE: Quantitative C-score			6.1	1.9
Linguistic C-score			6.2	2.0
Total C-score			6.3	1.8
Freshman GPA			2.3	.7

As with other Wechsler scales, the deviation I.Q.'s for any age group in the standardization sample of the *WAIS* were arbitrarily given a mean value of 100, and a standard deviation of 15. The *WAIS* means for the high school sample were approximately .5 σ above the *WAIS* standardization sample mean. For the college sample, the *WAIS* means were approximately 1.0 σ above the standardization sample means. These high school mean I.Q. values are comparable to those reported as typical for high school graduates (Cronbach, 1960), and the college freshman I.Q. values are similar to those reported by others (Cronbach, 1960; Plant, 1958; Plant and Richardson, 1958; Wall *et al.*, 1962). The *WAIS* I.Q.'s reported in Table 1 are evidence of the selectivity of formal educational attainment.

The predictive validity coefficients obtained for the high school and the college samples for all predictors are found in Table 2.

For each of the 17 *WAIS* predictors, the validity coefficients for the high school sample were higher than the corresponding coefficients for the college sample. In light of the differences between samples in terms of level and homogeneity of *WAIS* I.Q.'s, such differences in obtained coefficients are to be expected. It will be noted

TABLE 2

Predictor-Criterion Correlation Coefficients for the High School and College Samples

Predictor		H.S. r_{16}	Coll. r_{16}
WAIS: V subtests (Scaled scores)	Information	.54	.48
	Comprehension	.55	.33
	Arithmetic	.45	.19
	Similarities	.50	.39
	Digit Span	.37	.04
	Vocabulary	.65	.46
WAIS: P subtests (Scaled scores)	Digit Symbol	.34	.15
	Pic. Completion	.33	.20
	Block Design	.29	.19
	Pic. Arrangement	.22	.07
	Obj. Assembly	.17	.12
	Verbal	.63	.47
WAIS: Σ Scaled Scores	Performance	.43	.24
	Full-scale	.62	.44
WAIS: I.Q.'s	Verbal	.63	.45
	Performance	.44	.24
	Full-Scale	.62	.43
Terman-McNemar I.Q.		.72	
Primary Mental Abilities		.64	
ACE: Quantitative C-score			.20
Linguistic C-score			.38
Total C-score			.38

that the *WAIS* Verbal subtest scores yielded higher coefficients than did the Performance subtest scores for the high school sample, and generally for the college sample too. The *WAIS* Verbal composite scores correlated higher with the criteria than did the Performance composite scores.

In terms of correlation coefficient magnitude, the *WAIS* vocabulary subtest scaled score was as good a predictor of the criterion as any composite *WAIS* measure. This finding suggests that future attempts to devise the best short form of the *WAIS* might most profitably involve subtest prediction of meaningful non-test criteria. There may be several different best *WAIS* short forms, depending upon what is meant by best: best for the prediction of what criterion? To date, the criterion in studies of the best short form of the *WAIS* has been a *WAIS* composite score, and usually this has been the Full-Scale I.Q.

To determine whether criteria of academic success were as well predicted from *WAIS* scores as from group academic aptitude test scores, tests of the significance of the differences between correlated

test-criterion coefficients were undertaken for selected predictors. Table 3 summarizes the results of such tests.

TABLE 3

Significance of the Differences between Correlated Test-Criterion Coefficients for Selected Predictors

Sample	Variates	r_{12}	t
H.S.	WAIS Vocab. Scaled	.65	
	Terman-McNemar I.Q.	.72	.72
	WAIS Verbal I.Q.	.63	
	Terman-McNemar I.Q.	.72	.78
Coll.	WAIS Info. scaled	.48	
	ACE L C-score	.38	1.98*
	WAIS Verbal scaled	.47	
	ACE Total C-score	.38	1.84

* $p < .05$.

One of the four t ratios reported yielded a difference at or beyond the .05 level of significance, and this was for the highest WAIS sub-test validity coefficient compared with the highest group test validity coefficient in the college sample. In essence, these results are similar to those reported earlier in the two studies wherein individually administered vs. group administered ability tests were compared in terms of goodness of prediction of the same validity criterion.

It would appear that the level of accuracy in predictions of educational success from the WAIS or portions of the WAIS would be as high as would the level for the same predictions from a group test for the same criterion.

REFERENCES

- Conry, R. "The Prediction of Educational Success in High School and in College with the Wechsler Adult Intelligence Scale (WAIS) and Group Tests." Unpublished B.A. Honors Thesis, San Jose State Coll., 1963.
- Cronbach, L. J. *Essentials of Psychological Testing*. (2nd. ed.) New York: Harper, 1960.
- Guilford, J. P. *Fundamental Statistics in Psychology and Education*. (3rd. ed.) New York: McGraw-Hill, 1956.
- Plant, W. T. "Mental Ability Scores of Freshmen in a California State College." *California Journal of Educational Research*, IX (1958), 72-73.
- Plant, W. T. and Lynd, Celia. "A Validity Study and a College

- Freshman Norm Group for the WAIS." *Personnel and Guidance Journal*, XXXVII (1959), 578-580.
- Plant, W. T. and Richardson, H. "The I.Q. of the Average College Student." *Journal of Counseling Psychology*, V (1958), 229-231.
- Terman, L. M. and McNemar, Q. *Terman-McNemar Test of Mental Ability: Manual*. New York: World Book, 1941.
- Thurstone, L. L. and Thurstone, Thelma G. *American Council on Education Psychological Examination for College Freshmen: Manual*. Princeton: Educational Testing Service, 1948.
- Thurstone, L. L. and Thurstone, Thelma G. *Chicago Test of Primary Mental Abilities: Manual*. Chicago: Science Research Associates, 1953.
- Walker, Helen M. and Lev, J. *Statistical Inference*. New York: Henry Holt, 1953.
- Wall, H. W., Marks, L., Ford, D. H., and Ziegler, M. L. "Estimates of the Concurrent Validity of the WAIS and Normative Distributions for College Freshmen." *Personnel and Guidance Journal*, XXXX (1962), 717-722.
- Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: The Psychological Corporation, 1955.
- Wechsler, D. *The Measurement and Appraisal of Adult Intelligence*. (4th. ed.) Baltimore: Williams & Wilkins, 1958.

THE INTERACTION OF ABILITY LEVELS AND SOCIOECONOMIC VARIABLES IN THE PREDICTION OF COLLEGE DROPOUTS AND GRADE ACHIEVEMENT¹

BEN BARGER AND EVERETTE HALL

University of Florida

A continuing problem for educators has been the student who, although at least minimally qualified to complete his education, has either failed in his course work, or dropped out of school during the academic year.

Many investigators have studied various facets of the problem, and a large number of these studies have been summarized in *The American College* (Sanford, 1962). In general, these studies have treated the dropout as a unitary phenomenon, and have studied the relation of single variables to it (Hopkins *et al.*, 1958). An exception to this is a study in which a prediction index was developed from a biographical inventory (Aiken, 1963).

The Problem

The purpose of the present study was to determine: (1) whether the relationship of socioeconomic variables to dropping out of college is the same for different ability levels; and (2) whether there is a relationship between these same variables and grade achievement, when ability is controlled.

Method

All freshmen and sophomores entering the University of Florida for the first time in September 1961, and for whom background data

¹ Supported by NIMH Project Grant MH-380, the purpose of which is to develop, in a university setting, a comprehensive mental health program based on public health principles.

and *School and College Ability Test (SCAT)* scores were available, were used as the base population.

The entire sample (2348 males and 1296 females) was divided, males and females separately, into three groups of approximately equal size. The resulting groups represent, for each sex, the upper, middle, and lower thirds of ability (*SCAT*) for this sample.

The three ability groups were then compared, by chi-square, in terms of their distributions for the following variables: parents' marital status, parents' income, father's education, father's occupation, student's religious preference, ordinal position, and family size.

Then, within each ability level, those students who failed to complete the academic year (600 males and 309 females) were compared by use of chi-square, on each variable, to the students who completed the year.

These two steps provide a measure of the relationship of each variable to ability, as well as a comparison for different ability levels, of the relationship of these variables to dropping out of college.

As a next step, those students who completed the academic year were divided into three equal groups on the basis of their *SCAT* scores. Their distributions for each socioeconomic variable were then compared by use of chi-square.

For each of the three ability groups, another division into three equal groups was made on the basis of grade point average (GPA) for the academic year. Within each third of the ability distribution, then, there are three GPA groups which have only slight differences in mean ability scores. This made possible a comparison, within ability levels, of the ways in which differential achievement relates to the variables studied, with ability controlled.²

Results

The analysis of the results is based upon the data presented in Tables 1, 2, and 3. Reference to these tables may serve to clarify several of the findings to be reported.

² Data consisting of Tables A-Z and AA-JJ has been deposited with the ADI Publications Project. Order Document No. 8340, remitting \$2.25 for 35-mm. microfilm or \$5.00 for photoprints. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1

*Relationships between Background Variables and
Three School and College Ability Test Levels*

	Male			Female		
	χ^2	df	p	χ^2	df	p
Parents' Marital Status			NS			NS
Family Income			NS			NS
Father's Education	35.1	14	$<.01$	28.2	14	$<.02$
Father's Occupation	28.4	14	$<.02$	22.5	14	$<.10$
Religious Preference			NS	32.8	16	$<.01$
Ordinal Position and Family Size	33.3	18	$<.02$	33.9	18	$<.02$

TABLE 2

*Comparison of Dropouts against Non-Drops
on Background Variables by Ability Levels**

	Upper Third SCAT					
	Male			Female		
	χ^2	df	p	χ^2	df	p
Parents' Marital Status	9.2	2	$=.01$	27.9	2	$<.001$
Family Income	9.3	5	$<.10$			NS
Father's Education			NS	12.6	7	$<.10$
Father's Occupation	13.0	7	$<.10$	15.2**	7	$<.05$
Religious Preference			NS	13.7	8	$<.10$
Ordinal Position and Family Size			NS			NS

*No significant relationships existed for the lower and middle ability thirds for either males or females.

**May be unreliable; several expected frequencies less than 5.

Entering Students and Dropouts

Males. For all entering males, only father's education ($p < .01$), father's occupation ($p < .02$), and family size and position ($p < .02$) are related to ability level, when the procedure outlined above is used. There is a positive relationship between the amount of father's schooling and ability, but between father's occupation and ability there is a curvilinear relationship. That is, those students whose fathers are in the middle occupational levels have lower mean ability scores, on the average, than those students whose fathers come from the upper and lower occupational levels. For family size and ordinal position, the largest differences from the expected frequencies are for the oldest child in the family, especially

TABLE 3

*Relationships between Background Variables, Ability Levels, and Grades Achieved**

	Females								
	Lower Third SCAT			Middle Third SCAT			Upper Third SCAT		
	x^2	df	p	x^2	df	p	x^2	df	p
Parents' Marital Status	15.9	4	<.01			NS			NS
Family Income			NS	28.0	10	<.01	17.7	10	<.10
Father's Education			NS	25.2	14	<.05	30.0	14	<.01
Father's Occupation			NS			NS	18.1	10	<.10
Religious Preference			NS			NS			NS
Ordinal Position and Family Size			NS	32.2	16	<.01			NS

*No significant relationships existed for men.

for those males from families of four or more, in the direction of higher ability.

The only variable which relates to dropouts, for the males, when they have been divided into ability groups is current marital status of parents. For males in the upper third of ability only, those from broken homes (parent(s) deceased or divorced) tend to drop out at a higher rate than those males from intact homes ($p = .01$).

Females. For all entering females, religion ($p < .01$), father's education ($p < .02$), and family size and position ($p < .02$) are related to ability levels. For father's occupation, the probability level is only .10. Of the religious groups, the largest difference from the expected frequency is for the Jewish girls, who tend to have, on the average, lower ability scores in the sample studied. The Episcopal girls, and girls from minor Protestant sects, tend to have somewhat higher ability scores than the average for the sample.

There is a linear positive relationship between father's education and ability, except for the group whose fathers had three years of college. These girls show lower ability scores, on the average, than girls whose fathers had only one or two years of college. There is an interaction between family size and ordinal position in their relationship to ability scores. For example, the middle girls in a family of three, have lower ability scores, on the average, than expected. The middle girls, in families of four or more, have mean ability scores which differ little from the expected frequencies.

As for the males, when the sample is divided into thirds accord-

ing to ability, only parents' marital status has a significant association with dropping out of school ($p < .001$), and only for those girls in the upper third of ability. There is a tendency, also, for the girls from broken homes to drop out of college at a higher rate than those girls from intact homes.

Students Completing the Academic Year

Males. For those who completed the academic year, father's education ($p < .01$), and father's occupation ($p < .05$) are the only variables studied which are significantly associated with ability scores. The relationships for the two variables are approximately the same as for those discussed for the total sample above. This result would seem to confirm the finding above that there is no relationship between these variables and dropping out of college, since removal of the dropout sample does not alter the relationships shown.

When the males are divided by grades into three groups within each ability level (such that ability is controlled to a great extent) and when the distributions for each variable are compared, there are no significant differences. This finding would indicate that, for this sample, achievement differences over categories for a given variable (e.g., father's education) can be largely accounted for by differences in ability.

Females. For girls who completed the academic year, religion, father's education, father's occupation, ordinal position, and family size are all related to ability at the .05 level or higher. The relationships, within variables, are approximately the same as those described above for the total sample. Since the relationship for father's occupation is strengthened for this sample, there is probably a very slight relationship of this variable to dropping out, although the relationship does not reach significance with this method of comparison. The distribution for this variable is similar to that described for the males; that is, it is curvilinear.

When the females are divided into three grade achievement groups within ability levels, the following relationships are significant: marital status of parents for the lower third in ability ($p < .01$), family income for the middle third in ability ($p < .01$), father's education for the middle ($p < .05$) and high thirds ($p <$

.01) in ability, and ordinal position and family size for the middle third in ability ($p < .01$).

The girls from broken homes, in the lower ability group here, tend to achieve higher grades, on the average, than girls from intact homes. In the middle third of ability, the girls from the lower income levels (less than \$5,000, and \$5,000-\$8,000) tend to achieve higher grades, on the average, than girls from the higher income levels. For the higher ability third, girls whose fathers have college degrees or graduate training achieve higher grades, on the average, than those girls whose fathers had fewer years of schooling. For the middle third of ability, however, the relationships are reversed, but the differences are not so large.

For the middle third in ability, the younger girls in two-child families show a higher than average grade achievement, and middle girls, especially in the larger families, show a lower grade average than expected.

Discussion

These data indicate that, when one controls for ability, dropping out of college is not related significantly for this sample to any of the socioeconomic variables studied, with the exception of parents' marital status. Thus, studies which have shown such relationships without taking ability into account may have been simply reporting findings which resulted from the relationships of ability to the several socioeconomic variables studied.

Since parents' marital status is also related to grade achievement for females in the lower third of ability, one may infer that there is an interaction between parents' marital status, ability, and adaptation to the college environment. That is, among high ability students, who often do not develop effective study habits, those from broken homes may more frequently be unable to adapt to the sharply increased demands which students frequently experience in the transition from high school to college. On the other hand, lower ability females from broken homes may represent a more select sample in terms of academic motivation and adaptability than is represented by the lower ability students from intact homes.

The fact that none of the socioeconomic variables predicts grade achievement for the males is rather surprising, especially in view of

the several positive findings for females. Since the correlations between ability scores and grades are higher for the females than for the males, it was felt that objective nonintellective factors such as family income would explain at least some of the variance in academic achievement for the males. Again, differential selectivity related to various categories of the variables may alter relationships obtaining in the general population.

However, an alternative possibility is that social and family background factors affect a girl's perception of herself, and consequently her happiness at college, in a more direct way than they do for boys. Possibly a boy is accepted by peers more easily for what he can do, and therefore his background may be of less importance, than for a girl. In any case, girls of middle ability from the lower income levels seem to achieve better academically, on the average, than girls from higher income levels.

Father's education seems to be related to academic motivation for girls whose ability is commensurate with the level of their father's educational achievement. That is, high ability girls whose fathers are college graduates achieve higher grades than high ability girls whose fathers did not attend or complete college. On the other hand, middle ability girls whose fathers did *not* complete college achieve higher grades than middle ability girls whose fathers are college graduates.

Conclusions

- (1) In studying the relationships between socioeconomic variables and dropping out of college, investigators should employ some method of controlling for ability.
- (2) There is a strong relationship between some socioeconomic factors and ability in the general population, and these relationships may be altered sharply, by selection factors, in college populations.
- (3) At the University of Florida, knowledge of socioeconomic background adds to the prediction from ability scores of grades achieved in college by females. For males, however, when ability is controlled, there are no significant relationships between the socioeconomic variables studied, and grades achieved.
- (4) Various socioeconomic variables may have considerably dif-

ferent meaning psychologically for males and females who come to college.

REFERENCES

- Aiken, L. R. "The Prediction of Academic Success and Attrition from a Multiple Choice Biographical Inventory." Paper read at the Southeastern Psychological Association Meeting, Miami Beach, 1963.
- Hopkins, J., Malleon, N., and Sarnoff, I. "Some Non-intellectual Correlates of Success and Failure among University Students." *British Journal of Educational Psychology*, XXVIII (1958), 25-36.
- Sanford, R. N., (editor). *The American College*. New York: John Wiley and Sons, 1962.

NEED PATTERNS AND ABILITIES OF COLLEGE DROPOUTS

JAY L. CHAMBERS

Kentucky State Hospital, Danville, Kentucky

BEN BARGER

University of Florida

AND

LEWIS R. LIEBERMAN

Charles L. Mix Memorial Fund, Inc. Americus, Georgia

SCHOOL dropout is now recognized as a national problem, basic to the social and economic well being of the nation. Concern over dropouts has stimulated considerable research and has generated several programs designed to correct the problem. A list of such programs as well as a list of general references to school dropout studies has been prepared by the National Education Association Research Division (1963). From the data obtained thus far, the school dropout problem appears almost as complex as that of mental illness, and, indeed, the two are not unrelated. The dropout has failed to adequately cope with one of the most important demands of the American culture: formal education.

The present study attempted to analyze factors involved in college freshman dropout. It is a continuation of previous research reported to the 1961 Georgia Association of Junior Colleges (Chambers and Lieberman, 1961) and in other studies reported by Barger and Hall (1964a, 1964b). The first of these reports described motivation factors related to college adjustment and achievement of male students from two successive entering freshman classes at Georgia Southwestern College. Dropouts from these classes were successfully discriminated from survivors by use of the Picture Identification Test

(*PIT*). Based on differentiating *PIT* signs obtained from the study, the male dropout at Georgia Southwestern was described as admiring those who are aggressive, autonomous, sexually active, and defensive. He least liked those who are cautious, dependent, deferent, and nurturant toward others. His judgment was particularly poor for the defence and achievement needs.

The Picture Identification Test

The *PIT* has been modified since the above analyses were made. A description of the *PIT* form used in the present study is available in the literature (Lieberman and Chambers, 1963). Briefly, the *PIT* measures judgments, attitudes, and associations pertaining to 21 needs of the Murray (1953) need system. The technique of measurement is projective in that the subject (*S*) matches head and shoulder photographs of people of the same sex as *S*, with descriptions of the Murray needs. He also makes affective (liked-best and liked-least) selections of the photographs.

The Judgment (*J*) score for a need is based on the percentages of those in a normative group who made the same need-picture matchings as *S*. The Attitude (*Att*) score for a need is based on *S*'s preferences for either best-liked or least-liked photographs matched with the need. The Association Index (*AI*) for a need is based on the degree to which *S* conforms to a normative group in associating the need with other needs by attributing them to the same photograph. In addition, Sum *J* and Sum *AI* scores provide over-all measures of judgment and associative conformance. A measure of the over-all tendency to use affectively chosen photographs rather than neutral ones with which need descriptions are matched is measured by the Variability (*Var*) score. A high Consistency (*Con*) score indicates that *S* matched needs with more liked-best than with liked-least pictures, whereas a low *Con* score indicates the reverse tendency.

All scoring procedures for the *PIT* are objectively defined so that the test can be scored by electronic data processing equipment and computers.

The following interpretations of the three types of scores are offered: the *J* score for a need reflects *S*'s ability to judge when that need is appropriate according to the press of circumstances; the *Att* score indicates whether he feels it is better or worse than other

needs; the *AI* provides an indication of how well *S* is able to express or to satisfy the need when it is aroused.

The present study sought to confirm the generality and the reliability of the dropout need factors previously reported. In addition, the study was designed to investigate intellectual as well as motivation factors in college dropout.

Method

Subjects. All *Ss* were beginning freshman students at the University of Florida. A dropout was defined as any student who left the University for any reason during the fall or winter trimesters or who failed to return for the winter trimester for any reason. A comparative sample was selected from among students who finished both trimesters. A sample of 319 male dropouts was matched by an equal number of male survivors. A sample of 189 female dropouts was matched by an equal number of female survivors.

Procedure. During freshman orientation week preceding the fall trimester, all beginning freshman students were given the *PIT*, the *School and College Ability Test (SCAT)*, and a questionnaire which contained items concerning study habits and grade expectations. The *PIT* data were scored through the facilities of the University Computing Center and *SCAT* scores were obtained from the office of the Registrar with the approval of the Board of Examiners. All norms for the *PIT* scores were based on 1000 male and 1000 female freshman students at the University of Florida. At the end of the year, dropouts and survivors were selected from a sample of 1753 males and 1091 females for whom all the required data were available.

Analysis. Male and female data were analyzed separately and were subjected to discriminant function analyses processed by a computer. Because of the large number of *PIT* variables involved (67), the *J*, *Att*, and *AI* dimensions were first separately analyzed to cull those variables showing least promise for discriminating dropouts from survivors. The chosen *PIT* variables were combined with four other variables for final discriminant function analyses. The four non-*PIT* variables were *SCAT* Verbal and Quantitative scores, a self rating of grades expected, and a self rating of study effort which the student planned to exert. Thirty variables were

used in the final male discriminant function analysis and 37 variables were used in the final female analysis.

Results

Discriminant function analyses discriminated dropouts from survivors in both male and female groups ($F = 3.71$ for males and 2.35 for females; $p < .0001$ for both analyses). Among the males, 216 dropouts (68%) and 204 survivors (64%) were correctly classified by the discriminant scores. Among the females, 135 dropouts (71%) and 134 survivors (71%) were correctly classified.

TABLE 1
*The Distribution of Discriminant Scores by Deciles
for Survivors (S) and Dropouts (DO)*

Decile	Male		Female	
	S	DO	S	DO
10	55	9	31	7
9	43	21	29	9
8	39	25	28	10
7	41	23	22	16
6	33	30	24	13
5	30	33	20	17
4	26	38	14	24
3	31	33	6	32
2	7	57	9	29
1	14	50	6	32

The discriminant score distributions for male and female groups were broken into deciles to determine whether discrimination was better at the extremes of the scale than in the middle. Table 1 presents these decile distributions. Discrimination is obviously better at the extremes. If the upper and lower three deciles were considered as the extremes, 72% of the males were correctly classified in the extremes and 57% were correctly classified in the middle four deciles of the scale. For the females, 79% were correctly classified in the extremes and 58% were correctly classified in the middle of the scale.

Table 2 presents the ten variables which contributed most to D^2 for the male and female analyses. *SCAT V* and *SCAT Q* were the largest contributors for both analyses. High *SCAT* scores indicated survival as denoted by a "+" sign in the table. A positive *Att* score for *n* aggression was a contraindication for survival for both males

TABLE 2

*Variables with Highest Contributions to D² for
Survivor-Dropout Discriminant Function Analyses*

D ² Rank	Male		Female	
	Variable	Contribution to D ²	Variable	Contribution to D ²
1	SCAT V	.21 (+) ^a	SCAT V	.36 (+) ^a
2	SCAT Q	.17 (+)	SCAT Q	.21 (+)
3	Var Att	.06 (-) ^b	Def Att	.07 (+)
4	Agg Att	.05 (-)	Aut Att	.05 (-) ^b
5	Aff Att	.03 (-)	Agg Att	.05 (-)
6	Sex AI	.03 (+)	Aff J	.04 (+)
7	Dom Att	.02 (+)	Aff Att	.02 (+)
8	Rej Att	.02 (+)	Sex J	.02 (+)
9	Def J	.02 (+)	Har J	.02 (+)
10	Sum J	.02 (+)	Nur J	.02 (+)

^aHigh or positive score predicts survival.

^bLow or negative score predicts survival.

and females. The sexes were opposite with regard to *n* affiliation in that a positive *Att* score indicated survival for females but was a contraindication of survival among males. No other variables among the highest 10 contributors to D² were common to both sexes. The self rating variables did not appear among the 10 highest D² contributors for either sex. In general, the *PIT* results of the present study supported the findings previously reported by Chambers and Lieberman (1961).

Discussion

Intellectual factors (*SCAT V* and *Q*) accounted for approximately two thirds of the total contributions to D² in both male and female discriminant functions. The combined *PIT* need measures accounted for the remaining third (app.) of the D² contributions. *SCAT V* received a heavier weighting than *Q*, especially for females. The importance of verbal ability for survival in college among females may be due to curricular preferences, since women are less apt to enroll in *Q* type courses such as engineering and mathematics.

Need characteristics of dropouts and survivors can be theoretically interpreted from the *PIT* variables presented in Table 2. One could infer, for example, that the male dropout dealt with in this study likes to be aggressive and sociable (pos. *Att* scores toward *n*

aggression and n affiliation), but he does not like to assume leadership responsibilities and he finds it difficult to resist requests or demands from others (neg. *Att* scores for n dominance and n rejection). He tends to let his feelings become involved with his judgment (high Variability score). His approach to women is apt to be ineffective (low *AI* for n sex). He shows a general lack of perceptiveness concerning the requirements and demands made on him by circumstances (low Sum *J* scores), and he is particularly apt to misperceive situations with regard to the need to accept guidance and direction from others (low *J* score for n deference).

Again interpreting from Table 2, one might describe the female dropout as tending to be antisocial (neg. *Att* toward n affiliation), even when the situation calls for being friendly (low *J* scores for n affiliation). She likes to be aggressive (pos. *Att* toward n aggression). She dislikes control or discipline (pos. *Att* toward n autonomy), and she is antagonistic toward authority (neg. *Att* toward n deference). She is apt to display poor judgment with regard to sex, personal danger, and consideration for others (low *J* scores for n sex, and n harmavoidance, and n nurturance).

The one contradiction in the need makeup of male and female dropouts was a difference in attitude toward n affiliation. The results indicated that the male dropout likes to be sociable, whereas the female dropout tends to have an antisocial attitude.

As in the Georgia Southwestern dropout study cited in the introduction to this paper and in other studies (Michigan Department of Public Instruction, 1960; Greene, 1962; and Cassel and Coleman, 1962), a readiness to assert aggression appears generally characteristic of dropouts. These findings are consistent with the commonly held view that the prevailing mood of the dropout's opposite number, the scholar, is one of quiet, patient labor and contemplation rather than aggressiveness and destructiveness. They are also consistent with the fact that n aggression and n understanding are opposed needs according to the normal subjects who established the *PIT* Association Index.

Dropout prediction in the present study was not sufficiently accurate to warrant exclusion from admission to college of students with high dropout discriminant scores. However, the prediction was sufficiently accurate, especially for those with extreme scores, to sug-

gest an experimental program to try to change undesirable attitudes, perceptions, and associations of potential dropouts in an effort to prevent their withdrawal. In evaluating the prediction of dropouts in the present study, it is well to bear in mind that the dropout criteria used did not take into account the reasons for withdrawal. Thus, the result may have been obscured by instances where dropout was due to circumstances and hardships not attributable to the abilities and characteristics of the student.

Summary

Discriminant function analyses were separately applied to groups of 319 male and 189 female college freshman dropouts and survivors. Data analyzed consisted of *SCAT V* and *Q* scores; Picture Identification Test measures of need attitudes, judgments, and associations; and student self ratings of anticipated grades and study efforts. Discriminant function analyses differentiated male and female dropouts from their counterpart survivors ($p < .0001$ for both analyses). As expected, high *SCAT V* and *Q* scores predicted survival. According to the *PIT* discriminators, dropouts tended to be more aggressive, more resistant to authority and control, and had more problems with sexual adjustment than survivors.

REFERENCES

- Barger, Ben and Hall, Everette. "Need Ranks among Male College Students with Differing MMPI Profiles." Unpublished paper presented at the SEPA Meeting, Gatlinburg, Tenn., (April, 1964). (a)
- Barger, Ben and Hall, Everette. "Personality Patterns and Achievement in College." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIV (1964), 339-346. (b)
- Cassel, Russell N. and Coleman, Jack C. "A Critical Examination of the School Dropout, Reluctant Learner, and Abler Non-College Student Problem." *Bulletin of the National Association of Secondary School Principals*, XLVI, No. 277 (November, 1962), 60-65.
- Chambers, J. L. and Lieberman, L. R. "Personality Factors Relating to College Adjustment and Achievement." Unpublished paper presented at the Georgia Association of Junior Colleges, Annual Meeting, Georgia Southwestern College, 1961.
- Greene, Bert I. "Continuing Education and the High School Dropout." *Adult Education*, XII (1962), 76-83.
- Lieberman, L. R. and Chambers, J. L. "Differences between Prison-

- ers and Trade School Students on the Picture Identification Test." *Perceptual and Motor Skills*, XVII (1963), 355-361.
- Michigan Department of Public Instruction, State Curriculum Committee on Holding Power. "Quickie Kit on School Holding Power." Publication No. 507. Lansing: The Department, 1960.
- Murray, H. A. *Explorations in Personality*. New York: Oxford University Press, 1953.
- National Education Association, Research Division. "Project: School Dropouts." NEA Publications, Stock No. 732-18906, (April, 1963).

TWO CROSS VALIDATIONS OF THE OPINION, ATTITUDE AND INTEREST SURVEY

SAM C. WEBB
Emory University

THE *Opinion, Attitude and Interest Survey (OAIS)* is a paper and pencil inventory designed to assess several important personality and interest variables which seem especially relevant to the admission of college students and to their advisement during the first year. There are 14 scales based on empirically derived keys. The scales are grouped into four categories as follows: three response bias scales, three academic promise scales, three psychological adjustment scales, and five educational-vocational interest scales.

Purpose

This paper presents the results of two cross validations of this inventory based on data from students in the liberal arts college of Emory University. The first cross validation is based on data from a representative third ($N=205$) of the entering freshman class in the fall of 1962; the second is based on data from a representative third ($N=209$) of the entering freshman class in the fall of 1963. Each group included approximately 60 percent boys and 40 percent girls. All students were selected on the basis of a predicted average derived from a multiple regression equation involving the high school average (*HSA*) and the verbal (*SAT-V*) and mathematical (*SAT-M*) scores of the *Scholastic Aptitude Test* of the College Entrance Examination Board (Webb and McCall, 1953). The *OAIS* was administered to each group during orientation week. In the validation of the scales, an effort has been made to select criteria as nearly like those employed by the test author as circumstances would permit.

Results

Academic Promise Scales

The validity of the academic promise scales (first year college grade point average serving as a criterion), as well as intercorrelations, means, and standard deviations, are shown in Table 1. For comparative purposes similar data for the *HSA*, *SAT-V*, and *SAT-M* are included. In the computation of *HSA* and college grade point average, letter grades were assigned numerical values as follows: A = 40; B = 30; C = 20; D = 10; F = 0. For these variables the ordering with respect to validity for both years was *HSA*, Achiever Personality (*Ach P*), *SAT-V*, *SAT-M*, Intellectual Quality (*Int Q*), Creative Personality (*Cre P*). For the first three variables, all validities were significant at the .01 level.

To assess the contribution of the *Ach P* scale to the multivariate prediction of grades, the investigator computed a multiple validity coefficient involving use of *HSA*, *Sex*, *SAT-V* and *SAT-M* as predictors, both with and without *Ach P* as a fifth variable. The order of adding variables was determined by the Wherry test selection procedure (Garrett, 1957). When the *Ach P* scale was not considered, the order of including variables and of resulting shrunken cumulative *R*'s was *HSA* ($R=.56$); *SAT-V* ($R=.595$); *Sex* ($R=.604$); and *SAT-M* ($R=.611$) for the 1962 class. For the 1963 class the results were: *HSA* ($R=.52$); *SAT-V* ($R=.541$); *SAT-M* ($R=.542$). But when the *Ach P* scale was also considered, the results for the 1962 class were: *HSA* ($R=.56$); *Ach P* ($R=.603$); *SAT-V* ($R=.630$); *Sex* ($R=.637$); and *SAT-M* ($R=.644$). Also for the 1963 class the results were: *HSA* ($R=.52$); *SAT-V* ($R=.541$); *Ach P* ($R=.554$); *SAT-M* ($R=.560$). *Sex* did not enter the multiple correlation for this class.

According to the computation procedures of Fricke (1963) the percentage of improvement in variance accounted for when *Ach P* is used with the other variables is 11.09 for the 1962 class, and 6.74 for the 1963 class. (See Table 2.)

The percentage of improvement in variance accounted for by the use of *Ach P*, in addition to *HSA*, *SAT-V*, and *SAT-M*, for the sexes considered separately is shown in Table 2. These percentages of gain range from 9.09 to 25.27.

TABLE 1
Intercorrelations, Means, and Standard Deviations of College Grades and Six Predictors for 1982 (N = 205) and 1983 (N = 209) Classes

Variables	1	2	3	4	5	6	7	Mean	SD
1 Achiever Personality									
2 Intellectual Quality	-.03	-.03*	-.29	.10	.07	.38	.42	49	27
3 Creative Personality	-.19	.34	.25	.57	.19	.07	.20	62	27
4 SAT-Verbal	-.01	.42	.16	.19	-.02	-.22	-.12	49	31
5 SAT-Mathematical	-.16	.20	.04	.40	.33	.17	.30	556	87
6 High School Average	.32	.08	-.29	.22	.06	.09	.21	584	63
7 First Year College Grade Point Average	.28	.03	-.22	.27	.16	.52	.56	33	5
Mean	52	59	48	566	590	33	24	24	7
Standard Deviation	27	26	30	80	74	5	7		

*Data for 1983 above diagonal; * sign found at .05 level is equal to .14.

TABLE 2

Percentage of Gain in Variance Accounted for in R by Addition of Ach P to HSA, SAT-V, and SAT-M

Sex	Year	N	Shrunken R		Percentage of Gain
			without Ach P	with Ach P	
Women	1962	81	.550	.601	19.40
	1963	86	.517*	.540	9.09
Men	1962	139	.570	.616	16.81
	1963	123	.486	.544	25.27
Total	1962	205	.611	.644	11.09
	1963	209	.542	.560	6.74

*SAT-M did not enter multiple correlation.

These predictors were also validated relative to a criterion of drop-out at the end of two years for the 1962 class. Twenty-five percent of the class had dropped out. The point biserial correlations were: *Ach P* = .16; *Int Q* = .05; *Cre P* = .02; *HSA* = .16; *SAT-V* = .06; *SAT-M* = .04. The values for *Ach P* and *HSA* were significant at the .05 level.

The validity of the *Int Q* scale was also assessed by correlating it with scores on the SAT. As seen in Table 1, it correlated with *SAT-V* and *SAT-M* .57 and .19, respectively, for the 1962 class and .42 and .20, respectively, for the 1963 class.

The *Cre P* scale was validated against ratings of creativity (involving use of a five point scale) provided by four faculty members (one rater per student). Subjects were 24 students of the 1963 class enrolled in the college freshman humanities course. Work in this course consisted primarily of reading and discussing critically selected literary classics. In these ratings, creativity was defined as imagination and originality in thinking and as high capacity for organizing ideas. The validity of .43 was significant at the .05 level.

Psychological Adjustment Scales

For the 1962 class the Social Adjustment (*Soc A*) scale was validated against end-of-year dormitory counselor ratings (three point scale) of general social adjustment and leadership potential. The correlations were .12 for boys ($N=89$); .02 for girls ($N=46$); and .08 for the total group ($N=135$).

For the 1963 class, scores were validated against end-of-year

dormitory counselor ratings (made on a four-point scale) of social adjustment—defined as having good personal relations—and of leadership potential—defined as possessing traits indicating promise for becoming a group or campus leader. With use of these respective criteria, the validities of the *Soc A* scale were: .06 and .07 for boys ($N=78$, eight raters used, one rater per student); .34* and .19 for girls ($N=84$, average ratings of 11 raters calculated, one to three raters per student); and .12 and .11 for the total group of 162 students. When the two ratings were compared, correlation coefficients of .71* for boys, .76* for girls, and .75* for the total group were obtained.

The emotional adjustment scale (*Emo A*) was validated against end-of-year dormitory supervisor ratings (derived from a four-point scale) of emotional adjustment—defined as personality attributes associated with feelings of security, optimism, personal worth, and calmness. For the 1962 boys ($N=139$), the correlation coefficient was $-.17$. For the 1963 class, the correlation coefficient was $-.13$ for boys ($N=78$); $-.03$, for girls ($N=86$); and $-.14$, for the total group ($N=164$).

The masculine orientation (*Mas O*) scale was validated by computing the point biserial correlation between sex and raw score on this scale. The correlation was .76 for the 1962 class and .96 for the 1963 class.

Educational-Vocational Interest Scales

The educational-vocational interest scales were validated in terms of their ability to predict choice of major for the junior year. Three analyses based on the scores of 165 students of the 1962 class who declared a major at the time of registering for the junior year are presented. These students were grouped according to Fricke's (1963) classification of majors.

In the first analysis the average percentile rank on each interest scale was computed for each group. The results are shown in Table 3. These data can be interpreted in two ways. First, one may consider mean scores on the five scales for students in each declared major area (one reads across rows). There are only two areas—business and humanities—for which the highest mean score is on

* In this section correlations significant at .05 level of confidence are indicated by an asterisk.

TABLE 3

*Mean Percentile Ranks on Five Interest Scales for
Five Grouping of Declared Majors*

Declared Major Area	N	Interest Scales				
		Business	Humanities	Social Science	Physical Science	Biological Science
Business	11	49.0	39.6	47.6	19.8	17.1
Humanities	69	23.5	40.6	39.7	35.4	29.8
Social Science	35	34.0	36.0	35.7	35.6	31.9
Physical Science	32	32.2	25.4	46.0	41.7	33.0
Biological Science	18	33.6	44.1	45.7	38.6	38.1

the scale purporting to measure interest in that area. Secondly, one considers mean scores on any one scale across declared major areas (one reads down columns). There are only three scales—business, physical science, and biological science—for which the highest mean score is made by the students in the declared major area for which the scale was designed. Though not tested for significance, these data do not appear to offer much promise for making a reliable prediction of academic major.

As a second way of considering scores on these scales, the number of students in each declared major area having the highest score on each scale was determined. The results are shown in Table 4.

TABLE 4

*Number of Students in Each Declared Major Area
Whose Highest Percentile Rank Is in Each Measured Interest Area*

Declared Major Area	N	Business	Humanities	Social Science	Physical Science	Biological Science
Business	11	4	2	3	1	1
Humanities	69	6	23	12	17	11
Social Science	35	10	6	4	7	8
Physical Science	32	6	2	8	11	5
Biological Science	18	2	2	5	5	4

If independence among scales is assumed, the observed distribution of frequencies does not differ significantly from chance expectation ($\chi^2 = 24.62$, $df = 16$).

As a third way of analyzing these data, the number of students in each declared major area having the lowest ranking score on each scale was determined. The results are shown in Table 5. Again, if

TABLE 5

*Number of Students in Each Declared Major Area
Whose Lowest Percentile Rank Is in Each Measured Interest Area*

Declared Major Area	N	Business	Humanities	Social Science	Physical Science	Biological Science
Business	11	1	3	0	2	5
Humanities	69	16	14	7	9	23
Social Science	35	7	9	5	7	7
Physical Science	32	8	9	2	5	8
Biological Science	18	6	5	3	2	2

independence among scales is assumed, the observed distribution of frequencies does not differ significantly from chance expectation ($\chi^2 = 8.49$, $df = 16$). Thus these data seem not to support the validity of these scales for predicting academic majors.

Summary

Evidence to support the validity of the academic promise scales of the OAIS has been presented. In terms of the criteria used, the data available offer little support for the validity of either the psychological adjustment or the educational-vocational interest scales relative to prediction.

REFERENCES

- Fricke, B. G. *Opinion, Attitude and Interest Survey Handbook*. Ann Arbor, Mich.: Evaluation and Examination Division of the University of Michigan, 1963.
- Garrett, H. E. *Statistics in Psychology and Education*, (Fourth Edition). New York: Longmans, Green and Co., Inc., 1957.
- Webb, S. C. and McCall, J. N. "Predictors of Freshman Grades in a Southern University." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIII (1953), 660-663.

VALIDATION OF A CAREFULNESS TEST BATTERY¹

H. G. OSBURN

University of Houston

CECIL J. MULLINS

Personnel Laboratory, Wright Air Development Center,
Air Research and Development Command, United States Air Force

AND

DANIEL E. SHEER

University of Houston

THERE is considerable interest in the identification of factors that predict performance criteria, but which measure attributes other than those identified by conventional aptitude tests. One class of predictors that shows promise in this respect is made up of tests which attempt to measure the degree of caution or carefulness exercised in the solution of simple problems.

Carefulness may be thought of as a general trait descriptive of the approach or method of attack used in problem solving situations, e.g., the characteristic way an individual works with problems. One individual may proceed haphazardly with little deliberateness and checking, while another may proceed with caution, order, and precision or, in short, with care. Carefulness, thus, may be reflected in a student's school performance as related more to how systematically and organizedly he works and less to his knowledge of specific content areas.

In his factor-analysis of perceptual tests Thurstone (1938) identified a carefulness factor that was defined by loadings on tests requiring accuracy on meticulous, detail tasks. French (1951) re-

¹ This investigation was supported in whole by Contract AF 41(657)-409, Personnel Laboratory, Wright Air Development Center, Air Research and Development Command, United States Air Force.

ported a carefulness factor obtained by Guilford (1947), which was defined by factor loadings on the wrong scores of the following speeded tests: Directional Plotting, Complex Scale Reading, and Plotting. In another study reported by French, the carefulness factor was identified by factor loadings on the wrong scores of the same speeded tests and the percent right scores on the following tests: Highest Number (circle the highest number in lists of numbers) and Scattered X's (circle every x in pages of pied letters). French described the carefulness factor as the ability to perform a task accurately in spite of working for speed.

In a factor-analytic study by Zachert and Ivens (1952), the following tests, given under speeded conditions, were included in the test battery: Plotting Flexibility, Plotting, Number Symbol Flexibility, Answer Sheet Marking, and Estimation of Length. The rights scores of the first four tests loaded on what was described as a speed factor, and the wrongs scores of the same tests loaded on a carefulness factor. The evidence for the distinction between these two factors was based on the additional loadings of Dial and Table Reading and Speed of Identification on the speed factor and the low correlations between rights and wrongs scores among the different tests.

In an unpublished study, Cox (1959) included Broken Pattern, Score Checking, Clerical Carefulness, and Symbol Finding given under non-speeded conditions in a test battery which was factor analyzed. In one analysis (analysis I) these tests were scored percent right, and in another (analysis II) they were scored rights. In both analyses the last three tests loaded on a carefulness factor. In analysis I they loaded with the wrongs score on Plotting Flexibility given under speeded conditions. In analysis II they loaded with the rights score of Counting Accuracy given under power conditions. Thus it appears that carefulness factors can be measured by simple detail tests administered under speeded conditions with respect to which wrong scores may be used and also by similar tests administered under power conditions relative to which rights scores are obtained. The relationship between these two factors is not clear.

The above studies have been concerned with the identification of a possible carefulness factor defined by the wrongs score or by the percent right score on highly speeded detail-type tests or by the rights score on unspeeded detail tests. The evidence from these studies suggests that identifiable factors do exist in this domain;

therefore, a logical line of further inquiry concerns the validity of such tests in the prediction of performance criteria. Do such tests contribute to the prediction of performance criteria over and above the variance accounted for by conventional aptitude tests alone? The purpose of the present study was to investigate the validity of a battery of carefulness tests against technical school criteria and to determine whether such tests do, in fact, increase the predictability of performance criteria.

Method

The subjects were 687 airmen enrolled in three technical courses at Shepard Air Force Base during 1961. The three courses, selected on the hypothesis that a carefulness component was involved in criterion performance, were (1) Communications Center Specialist, (2) Data Processing Machine Operator, and (3) Reciprocating Engine Mechanic.

The experimental battery, containing 11 relatively simple tests requiring attention to detail in order to avoid errors, was administered to all subjects prior to training. (The experimental battery consisted of 15 tests. Results on the four "risk" tests in the battery are not reported in this paper.) Two scores were obtained for each test: (1) a rights score and (2) either the number wrong or the percent right. Three of the tests—Score Checking, Number Symbol Flexibility, and Fact Inspection—were highly speeded. Four tests—Clerical Carefulness, Plotting, Estimation of Length, and Plotting Flexibility—were moderately speeded. The remaining tests—Letter Counting, Letter Matrix, Perceptual Discrimination, and Table Reading—were unspeeded. Additional test data obtained from Air Force records included: Armed Forces Qualifying Test (AFQT), Airman Qualifying Examination (AQE) Aptitude Indices: AQE-Mechanical, AQE-Administrative, AQE-General, AQE-Electronic, and Educational Level.

The criterion measures were: (1) Final Course Grade, (2) Performance Score, which represented the student's standing in the course as measured by performance tests exclusively, (3) Multiple Choice Score, which represented the student's standing in the course as measured by multiple choice verbal items exclusively, (4) Academic Washbacks, and (5) Carefulness Rating, obtained from the

instructors for each block of instruction. (An average of about six ratings per student was secured.)

Two scores were analyzed for each carefulness test with the exception of Fact Inspection for which only the conventional number correct score was used. For the other carefulness tests, a derived score, either the percent right or the number wrong, was analyzed in addition to the conventional rights score.

Results

The validity coefficients based on the number right score for the carefulness tests are reported in Table 1. The correlations are modest, although with over 200 cases in each sample over half are statistically significant at the .05 level. It is evident that the validities are slightly higher in the Communications Center Specialist sample than in the other two samples. This result was expected, since the communications course seemed to involve less content and more procedural detail than either of the other two courses. Table Reading, an unspeeded test, showed the highest validity of any of the carefulness tests, and as a group the unspeeded tests (Table Reading, Perceptual Discrimination, Letter Matrix, and Letter Counting) seemed to hold up better than the others.

The validity coefficients based upon the derived scores are presented in Table 2. It is evident that taken as a whole the correlations of the derived scores are somewhat lower than are the validities of the conventional rights score. The relatively unspeeded tests were least affected, since for these tests the derived score is just a linear transformation (to the extent that all subjects completed all items) of the conventional rights score. One highly speeded test, Score Checking, showed a consistent increase in validity for the derived score, whereas another highly speeded test, Number Symbol Flexibility, showed a marked decrease in validity for the derived score relative to the validity obtained through using the conventional rights score.

With regard to the possible differential validity of the tests in predicting the various criteria, it should be noted that there are only three experimentally independent criteria—Performance Score, Multiple Choice Score, and the Carefulness Rating. Final Course Grade and Academic Washback are both functions of the Performance and Multiple Choice Scores. It was expected that the Careful-

TABLE 2
Correlations between the Detail Carefulness Tests and Criterion Measures Using the Derived Scores (Decimal Points Omitted)

Predictor Tests	Final Course Grade			Academic Washbacks			Carefulness Rating			Multiple Choice Score			Performance Score		
	CCS ¹	DPS ²	RMS ³	CCS	DPS	RMS	CCS	DPS	RMS	CCS	DPS	RMS	CCS	DPS	RMS
Clerical Carefulness ⁴	08	05	11	08	00	13	14	-05	07	01	03	07	-02	07	13
Plotting ⁴	11	20	15	01	08	09	12	15	10	10	24	15	04	19	05
Estimation of Length ⁵	04	16	14	04	10	04	09	09	04	00	13	10	-04	14	13
Letter Counting ⁴	21	14	14	26	07	16	19	13	05	21	-07	07	26	17	15
Plotting Flexibility ⁵	06	12	11	-07	11	02	22	06	16	06	08	11	05	14	10
Score Checking ⁴	14	19	16	-01	16	19	13	21	10	08	06	07	08	21	22
Number Symbol Flexibility ⁵	02	06	-07	-09	01	00	08	05	-05	-01	01	-05	-03	07	-06
Letter Matrix ⁴	19	16	21	13	09	07	25	17	11	15	14	13	14	15	24
Perceptual Discrimination ⁴	17	15	14	16	10	06	09	11	03	09	06	13	17	16	15
Table Reading ⁴	33	26	22	23	27	14	30	22	16	31	22	13	32	28	24
Fact Inspection ⁶	01	08	18	04	20	-03	09	12	13	02	09	08	03	14	20

¹ CCS—Communications Center Specialist Sample ($N = 225$).

² DPS—Data Processing Specialist Sample ($N = 225$).

³ RMS—Reciprocating Engine Mechanic Sample ($N = 230$).

⁴ Scored Percent Right.

⁵ Scored Number Wrong and Reflected to Yield Positive Correlations.

⁶ Scored Number Right.

ness tests would correlate higher with the Performance Score and with the Carefulness Rating than with the Multiple Choice Score. As can be seen from both Tables 1 and 2, there is no marked trend in this direction.

As shown in Tables 1 and 2, the carefulness tests show low, but, in many cases, statistically significant validity coefficients. However, the *crucial question* is, of course, whether or not such tests account for criterion variance over and above the variance accounted for by conventional aptitude tests. It was observed that the carefulness tests were relatively independent of the operational aptitude tests, especially in the case of the derived scores. Therefore, it was possible that the carefulness tests taken as a group would add to the predictability of the various criteria. To investigate this possibility, the investigator computed the squared multiple correlation coefficient through use of all the available predictor variables, i.e., the six aptitude measures plus the derived scores on the 11 carefulness tests. This procedure is referred to as the full model. The full model was then compared to the restricted model in which the restricted model is the squared multiple correlation coefficient involving use of only the six aptitude measures. If the squared multiple correlation coefficient for the full model is significantly higher than that for the restricted model, one can conclude that the carefulness tests do contribute to the prediction of the criterion over and above the prediction afforded by the aptitude measures alone. The results of these analyses are presented in Table 3.

The data in Table 3 show that in the Communications Center Specialist sample the carefulness tests added significantly to the multiple correlation for four out of the five criteria. Generally speaking, the carefulness tests increased the prediction of those criteria that were rather poorly predicted by the aptitude measures alone. Only in the case of the Multiple Choice Score, which was well predicted by the aptitude measures, did the full model fail to reach significance.

The results are not so clear in the Data Processing Specialist sample as for the Communication Center Specialist samples. In this sample the multiple correlation coefficient was significantly increased for the Performance Score and the Academic Washback criteria in a manner consistent with the findings in the Communications Center Specialist sample. However, the full model failed to

TABLE 3
The Effects on Prediction of Adding the Eleven Carefulness Tests to the Aptitude Variables

	Communications Center Specialist ($N = 225$)				Data Processing Specialist ($N = 225$)				Reciprocating Engine Mechanic ($N = 230$)			
	$R^2_{(1)}$	$R^2_{(2)}$	$R^2_t - R^2_t$	F Ratio	R^2_t	R^2_t	$R^2_t - R^2_t$	F Ratio	R^2_t	R^2_t	$R^2_t - R^2_t$	F Ratio
Final Course Grade	.251	.160	.091	2.31 ²	.308	.263	.045	1.28	.295	.256	.039	1.06
Academic Washback	.276	.186	.090	2.37 ⁴	.191	.090	.101	2.43 ⁴	.214	.168	.046	1.11
Carefulness Rating	.213	.073	.140	3.38 ⁴	.326	.265	.061	1.77	.138	.096	.042	.93
Multiple Choice Score	.348	.293	.055	1.61	.321	.238	.083	2.39 ⁴	.257	.236	.021	.55
Performance Score	.308	.183	.125	3.42 ⁴	.278	.196	.082	2.22 ³	.272	.222	.050	1.32

(1) R^2_t —The squared multiple correlation coefficient relative to use of the six aptitude measures plus the eleven carefulness tests.

(2) R^2_t —The squared multiple correlation coefficient relative to use of only the six aptitude measures.

³ Significant at .05 level.

⁴ Significant at .01 level.

reach significance for the Final Course Grade and Carefulness Rating criteria. Moreover, the significant increase in the prediction of the multiple choice score did not lend itself to ready interpretation. One factor that might have been operating in this sample was the fact that both Final Course Grade and the Carefulness Rating were well predicted by the aptitude measures alone.

No significant results were obtained in the Reciprocating Engine Mechanic sample. It would appear that in this sample performance measures did not reflect a carefulness component.

Discussion

Two findings from this study are very suggestive. The major result was that the carefulness tests did function as expected in the Communications Center Specialist sample. It was this course that was judged, on the basis of task analysis, as most likely to involve a carefulness component in criterion performance. In general, this course involved the learning of fairly complex procedural systems that the student had to follow to the letter. On the other hand, the conceptual aspects of the course were relatively minor. The fact that in this sample the carefulness tests made an independent contribution to the prediction of all criteria except the Multiple Choice Score was exactly in line with expectation. These results must, of course, be tempered by the lack of positive results in the Reciprocating Engine Mechanic sample and the somewhat ambiguous results in the Data Processing Specialist sample. In the case of the Reciprocating Engine Mechanic sample, it is quite possible that a carefulness component is not involved in school criteria even though this aspect of performance may be quite important on the job.

One line of investigation suggested by this study concerns the measurement of carefulness. Our results seem to indicate that the most promising carefulness measures are unspeeded tests which involve a relatively extended stimulus-response chain. For example, Table Reading requires the subject to enter a first table to obtain data which are used to enter a second table to find the correct answer. Much further research is, of course, necessary to establish the carefulness domain as a possible source of hitherto unexplained criterion variance and to develop improved measures of carefulness factors.

REFERENCES

- Cox, J. A. "A Factor of Carefulness." Unpublished study, Air Force Personnel and Training Research Center, June, 1959.
- French, J. W. "The Description of Aptitude and Achievement Tests in Terms of Rotated Factors." *Psychometric Monographs*, No. 5, 1951.
- Guilford, J. P. (Ed.). *Printed Classification Tests*. Army Air Forces Aviation Psychology Program Research Report No. 5. Washington: Government Printing Office, 1947.
- Thurstone, L. L. "The Perceptual Factor." *Psychometrika*, III (1938), 1-17.
- Zachert, V. and Ivens, F. C. "Factor Analysis of the Airman Classification Battery AC-1B and the Texas Battery." Research Note Pers: 52-45. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, 1952.

AN EVALUATION OF AN ATTITUDE SCALE TOWARD TEACHING

PAUL L. CRAWFORD
Ohio University

RECENTLY, a self-administering teacher attitude scale, which can be taken with complete anonymity, was designed by Uhrbrock¹ (1962). The 50 item scale, which has a reported split-half reliability of .95, was based on the Thurstone (1959) scaling technique. The scale values of the test items, multiplied by 10, ranged from 12.6 to 98.5 with a neutral value of approximately 60.0.

Purpose

The present study was designed (1) to determine test-retest reliability of the scale, (2) to test for sex differences, and (3) to evaluate the relationship between student attitude scores and the course grades in an educational psychology course.

Methods and Results

In the first portion of the study, the Teacher Attitude Scale was administered to 342 educational psychology students, who were then retested 10 weeks later. The test-retest reliability coefficient for male students ($n = 121$) was .71; the test-retest reliability coefficient for female students ($n = 221$) was .65. The findings are about the same as Callis (1950) reported for the *Minnesota Teacher Attitude Inventory (MTAI)*, which is described by Cook, Leeds, and Callis (1951).

¹The teacher attitude scale by R. S. Uhrbrock has not been published commercially; however, anyone desiring a copy of the scale for research purposes may obtain a mimeographed copy from the author.

The mean score for the male students on the test in its initial administration was 62.7; the retest mean score was 64.3. The mean score for the female students on the first administration of the test was 67.3; the retest mean score was 69.2. The increases in scores by sex were not significantly different; however, female students scored significantly ($p < .01$) higher than male students on both the initial test and the retest.

In the second part of the study, the interrelationships among attitude scores (Variable X), numerical grade in educational psychology course (Variable Y), and *Ohio State University Psychological Test (OSUPT)* (Variable Z) were determined for a group of freshman students ($n=77$) as well as for a group of upperclassmen ($n=65$). These subjects were students from the first group for whom *OSUPT* scores were available. The following correlation coefficients were found for freshmen: $r_{xy} = +.18$, $r_{xz} = -.12$, $r_{yz} = +.45$; for upperclassmen: $r_{xy} = -.03$, $r_{xz} = -.01$, $r_{yz} = +.64$.

Discussion

Slight but nonsignificant increases between test and retest scores were found for both sexes, which may suggest increase in familiarity with the scale, or perhaps a slight change in attitude toward teaching as a result of the nature of the educational psychology course. Sandgren and Schmidt (1956) reported that practice teaching contributes substantially to favorable teacher attitude change. Female students scored significantly higher than male students on both test and retest—a result which is just the opposite of what Beamer and Ledbetter (1957) found when they used the *MTAI*. However, no correlation coefficient between attitude scores and grades or raw scores was significant; therefore, there appears to be no relationship between this attitude scale and the test performance in an educational psychology course, or between the attitude scale and a test of scholastic aptitude.

Summary

A teacher attitude scale was administered to 342 students enrolled in an educational psychology course and readministered 10 weeks later to determine the test-retest reliability of the scale. For male students the reliability coefficient was .71; for female students it was .65. No significant change in attitude scores was found for either

male or female students from the original to subsequent testing; although female students scored significantly ($p < .01$) higher than male students on both the test and the retest. Correlations between the attitude scale scores and the course grades, as well as between these scores and standing on the *Ohio State University Psychological Test* were nonsignificant.

REFERENCES

- Beamer, G. C. and Ledbetter, Elaine W. "The Relation between Teacher Attitudes and the Social Service Interest." *Journal of Educational Research*, L (1957), 655-666.
- Callis, R. "Change in Teacher-Pupil Attitudes Related to Training and Experience." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, X (1950), 718-727.
- Cook, W. W., Leeds, C. H., and Callis, R. *The Minnesota Teacher Attitude Inventory*. New York: Psychological Corporation, 1951.
- Sandgren, D. L. and Schmidt, L. G. "Does Practice Teaching Change Attitudes toward Teaching?" *Journal of Educational Research*, XLIX (1956), 673-680.
- Thurstone, L. L. *The Measurement of Values*. Chicago: University of Chicago Press, 1959.
- Uhrbrock, R. S. "Attitudes of Teachers." *Indian Journal of Psychology*, XXXIII (1962), 175-180.



MEASUREMENT OF ACHIEVEMENT MOTIVATION IN ARMY SECURITY AGENCY FOREIGN LANGUAGE CANDIDATES¹

WILLIAM E. DATEL²

Walter Reed Army Institute of Research

FORREST D. HALL³

Del Rey Oaks, California

AND

CHARLES P. RUFÉ⁴

Foreign Area Specialist Training Program

THE three achievement scales of Gough's (1957) *California Psychological Inventory* provide a means toward the assessment of the contribution afforded motivational, non-intellectual factors in academic achievement. Prior studies (Gough, 1961; Holland, 1959; Maxwell, 1960; Rosenberg, McHenry, Rosenberg, and Nichols, 1961) have demonstrated this fact. The data presented in this report represent an additional increment in the direction of delineat-

¹ While this study would not have been possible without the collaborative efforts of all three authors, the senior author takes sole responsibility for the statements made in this report. Special gratitude is extended to Consulting Psychologists Press, Inc., publishers of the *California Psychological Inventory*, for its kind permission to reproduce without obligation the three copyrighted scales for research purposes. The five minutes of electronic computer time needed to perform the innumerable computations in the data analysis is hardly representative of our appreciation to Professor Douglas G. Williams, United States Naval Post Graduate School, Monterey, California for his very major contribution to the computer programming aspect of the study.

² Formerly Clinical Psychologist, Mental Hygiene Consultation Service, Fort Ord, California.

³ Formerly Commanding Officer, Army Security Agency Personnel Procurement and Processing Detachment, Fort Ord, California.

⁴ Formerly Training Officer, United States Army Language School, Presidio of Monterey, California.

ing a measureable relationship between achievement motivation and academic success.

The study was originally conceived with the goal in mind of fortifying the selection procedure for entry into foreign language study at a United States Army speciality school. Though the results obtained fall short of this objective, it is felt that the findings presented are of interest when viewed from the perspective of the assessment of achievement motivation.

Method

Achievement via Independence (*Ai*), Intellectual Efficiency (*Ie*), and Achievement via Conformance (*Ac*) are the three scales of the *California Psychological Inventory (CPI)* which are thought to measure academic potential and achievement motive. The items comprising these three scales, 105 in all, were chosen from the total pool of 480 items making up the *CPI* and were mimeographed into test booklets headed by the usual "agree-disagree" instructions. The three scales were administered in this form to 290 incoming Army Security Agency soldier-students at the United States Army Language School (USALS)⁵, Presidio of Monterey, California in the fall of 1961. The students then went on to study any one of 18 different foreign languages, usually for a period of 47 weeks, though in a few instances the course was limited to six months.

Each of the soldiers in the study entered the Army through the Army Security Agency (ASA) recruitment procedures and enlisted with the avowed interest and desire to study a foreign language at USALS upon completion of basic training. In addition to the three *CPI* measures on each man, the following indices which were also obtained constitute an integral part of the data analysis in this study: age, years of education, score on the General Technical (*GT*) area from the Army Classification Battery, and Army Language Aptitude Test (*ALAT*) score. For every student the final grade received in the course and a dichotomous score of either successful completion or dropout status were also available.

Results

In Table 1 are the means and standard deviations for each variable on the total sample of 290 soldier-students.

⁵ Now called Defense Language Institute, West Coast Branch.

TABLE 1

*Means and Standard Deviations for Each of the Eight Variables
on the Total Sample Studied*

Variable	ASA Soldier-Students (N = 290)	
	Mean	Standard Deviation
Age	21.93	2.34
Education	14.61	1.68
GT	140.14	8.35
ALAT	40.51	7.32
Ai	23.63	4.68
Ie	41.75	3.81
Ac	28.35	4.67
Grade	84.56	7.66

In his interpretation of Table 1, the reader should be advised that the *GT* test was originally standardized with a mean of 100 and a standard deviation of 20. It is highly correlated with *Wechsler-Bellevue Intelligence Scale Form I* and *Wechsler Adult Intelligence Scale* IQ scores (Hedlund, 1959; Montague, Williams, Lubin, and Giesecking, 1957). The usual cutting score on the *ALAT* for a soldier to qualify for entrance into USALS is 30. In the case of students who successfully completed their course, the grade used was the final grade; in the case of dropouts the grade used was the last mark recorded for the student prior to his leaving the course.

Of the total sample of 290 students studied, 269 students successfully completed the course and 21 students were dropouts.⁶ This represents an attrition rate of 7.2 percent for the particular sample studied.

Table 2 presents a breakdown of completions versus dropouts on each of the eight variables. For each of the subsamples the means, standard deviations, differences between means, and critical ratio figures are given on each of the variables.

A multiple correlation analysis was also run on the raw data of this study in the hope that the establishment of regression coefficients for each of the variables studied would improve the efficiency of ASA language student selection. Unfortunately, the multiple correlation coefficient turned out to be of somewhat discouraging

⁶ The 21 dropouts can be further categorized into 10 academic failures and 11 administrative releases. There is oftentimes a very fine line between academic and administrative dropouts, so, for the purpose of this data analysis, the dropout samples were pooled.

TABLE 2

*Comparison between Completions and Dropouts
on Each of the Eight Variables Studied*

Variable	Completions (N = 269)		Dropouts (N = 21)		Difference between Means	Critical Ratio
	Mean	Standard Deviation	Mean	Standard Deviation		
Age	21.99	2.36	21.14	1.80	.85	2.03*
Education	14.63	1.70	14.33	1.39	.30	.92
GT	140.12	8.51	140.43	6.05	-.31	.22
ALAT	40.51	7.38	40.57	6.65	-.06	.04
Ai	23.79	4.76	21.67	3.06	2.12	2.92**
Ie	41.89	3.76	40.00	4.22	1.89	1.99*
Ac	28.51	4.55	26.29	5.81	2.22	1.72
Grade	85.51	6.48	72.33	10.71	13.18	5.56**

*Significant beyond .05 level.

**Significant beyond .01 level.

magnitude. Nevertheless, Table 3 is presented to the interested reader as a synopsis of the multiple correlation analysis.

For the entire sample of 290 students, Table 3 is a presentation of the correlation coefficients for six⁷ of the seven initial measures with grades as the criterion. The regression coefficients obtained are also listed for each of the independent variables, and the multiple correlation coefficient is stated.

TABLE 3

*Synopsis of Multiple Correlation Analysis of Data from Sample Studied:
Correlation of Grades with Each of Six Variables; Regression Coefficients
for Each of Six Variables; Multiple Correlation Coefficient*

Variable	ASA Soldier-Students (N = 290)	
	Correlation with Grades (r)	Regression Coefficient
Education	.15*	.27
GT	.19**	.06
ALAT	.32**	.28
Ai	.14*	.13
Ie	.12*	.07
Ac	.04	-.04
Multiple correlation coefficient:		.35**
Coefficient of determination:		.12

*Significant beyond .05 level.

**Significant beyond .01 level.

⁷ Age was deleted as a variable in the multiple correlation analysis. A "dry run" on the data revealed that the r for age with grades was .092.

Discussion

The sample studied is a highly selective one intellectually. The mean *GT* score for the group is over 140 (Table 1), which is two standard deviations above the standardized mean of 100! In view of the restriction of range associated with this upper extreme of intellectual talent, it perhaps makes somewhat more understandable the rather small mean differences found throughout the study (Table 2) as well as the very modest correlation coefficients in the grade correlation analysis (Table 3).

Because of this very restrictive spread of the sample, some attention can be paid to the positive findings which do occur in the data. It is extremely interesting, for example, that two of the achievement scales (*Ai* and *Ie*) succeed in significantly discriminating the "completers" from the "dropouts," whereas the more traditional predictors of academic achievement (i.e., the *GT* and the *ALAT*) fail to differ in the split sample (see Table 2). This finding may mean that given a certain *GT* and *ALAT* minimum or optimum score, the achievement scales can serve to sharpen the prediction of academic success. It must be admitted that this speculation could not be directly confirmed from the present data. But it can be said without conjecture that students who successfully complete a foreign language course do tend to obtain higher initial scores on *Ai* and *Ie* (and likewise on the *Ac*) than do students who for some reason or other fail to complete successfully the course. On the basis of these present data, the same statement cannot be made with respect to the *GT* and *ALAT* indices.

Turning to the results in Table 3 one notes that none of the correlations of test measures with grades gives cause for celebration, but it is of interest that five of the six *r*'s at least attain statistical significance. The multiple correlation coefficient of .35, with its coefficient of determination of only 12 percent, of course indicates that there is a large amount of unmeasured, error variance when grades are used as a criterion. Again, when the very restrictive spread of talent in the sample is considered, the findings appear to be reasonable.

Both the *ALAT* and the *GT* apparently correlate higher with grades than do any of the three achievement scales. Yet, as will be noted in Table 2, the *ALAT* and *GT* do not separate so adequately

the "completions" from the "dropouts" as do the *Ai* and the *Ie* scales. This discrepancy in the findings demands further discussion.

Could it be that the achievement scales, rather than measuring a linear, quantitative variable such as that represented by grades, might better be thought of as indicating simply the presence or absence of the achievement motive? Overly simplified, the data seem to say: When a candidate indicates that he has the motive to achieve (by scoring high on the achievement scales), he tends to complete successfully the course but does not necessarily receive an academic grade which is related in magnitude to his achievement scale scores. Furthermore, when a candidate receives a high score on the *ALAT* or *GT*, he tends to receive a high academic grade, though he is not necessarily by virtue of his high *ALAT* or *GT* score apt to complete successfully the course.

This explanation means, in effect, that the achievement scales add another, and albeit a rather more important other, dimension to the prediction of academic success than that rendered by mere aptitude or intellectual assessment. What appears at first glance as an inconsistency in the findings of this study actually emerges as more of a discovery than a discrepancy.

This line of thought gives rise to the proposal that increased attention be given toward the development of a selection battery for candidates entering army speciality schools which includes formal assessment of the achievement motive. The findings indicate that such measurement devices are already available in the form of scales like the *Ai*, *Ie*, and *Ac* of the *California Psychological Inventory*.

Summary

Three achievement scales from Gough's *California Psychological Inventory* were added to the traditional screening armamentarium prior to 290 Army Security Agency soldier-students' engagement in full time study of a foreign language. In terms of use of the difference-between-means method of analyzing the data on successful "completions" versus "dropouts," it was found that when the attrition rate for the sample was only 7.2 percent, two of the achievement scales significantly differentiated the two subsamples, whereas traditional measures (amount of formal education, *GT*, and *ALAT*) failed to do so. A multiple correlation analysis in which grades were used as the criterion was run on the data and, although yielding a

very modest multiple correlation coefficient, did produce results which, when viewed in conjunction with the difference between means analysis, suggest that a construct of achievement motivation emerges as a distinct dimension apart from aptitude or intelligence in the prediction of academic success. The findings provide encouragement for the incorporation of achievement motivation scales into selection batteries for academic performance.

REFERENCES

- Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- Gough, H. G. "A Comparison of College Dropouts and College Graduates." Paper read at the annual meetings of the American Personnel and Guidance Association, 1961.
- Hedlund, J. L. "Army Classification Scores: Their Meaning and Clinical Implication." Paper prepared for the Conference of Army Clinical Psychologists, sponsored by The Surgeon General, Department of the Army, Washington, D. C., September, 1959.
- Holland, J. L. "The Prediction of College Grades from the California Psychological Inventory and the Scholastic Aptitude Test." *Journal of Educational Psychology*, L (1959), 135-142.
- Maxwell, Martha J. "An Analysis of the California Psychological Inventory and the American Council on Education Psychological Test as Predictors of Success in Different College Curricula." *The American Psychologist*, XV (1960), 425. (Abstract)
- Montague, E. K., Williams, H. L., Lubin, A., and Giesecking, C. F. "Army Tests for Assessment of Intellectual Deficit." *United States Armed Forces Medical Journal*, VIII (1957), 883-892.
- Rosenberg, L. A., McHenry, T. B., Rosenberg, Anna M., and Nichols, R. C. "The California Psychological Inventory as a Potential Screening Device in an Academic Setting." Paper read at the annual meetings of the American Psychological Association, New York City, September, 1961.



AMERICAN COLLEGE TEST (ACT) PERFORMANCE AS A FUNCTION OF EXAMINEE ACCEPTANCE OF TEST

HENRY F. DIZNEY, ELINOR A. ELFNER, AND HORACE A. PAGE
Kent State University

Of all the forms of test validity in current use, face validity is generally held in lowest esteem by measurement specialists. The common theme in measurement literature equates face validity with public relations efforts designed to enhance the acceptability of the test in question (Anastasi, 1954; Cronbach, 1960a; Cureton, 1951). Mosier (1947) explicitly discussed "consumer acceptance" of the testing instrument as a type of face validity which he called "appearance." Cronbach (1960b) suggests the possible negative effects of low face validity upon test performance. A general impression gained from the scant literature available is that face validity in any form is, at best, a tolerable test characteristic to be considered secondary to empirical validation. Although one may not deny the crucial role of empirical validity, the possibility that the examinee's perception of the examination may affect his test performance, still exists. Goslin (1963) indicated that the effects of testing upon the examinee can have a bearing upon achievement by influencing self-confidence and self-perception. That these effects may be directly related to test performance seems to be a neglected issue.

As Sarason (1959) pointed out, it is rather amazing that in a highly test-conscious American culture, few systematic studies of examinee attitudes toward tests exist. It would seem, then, that studies designed to describe examinee test attitudes and to investigate the interaction effects of such attitudes on test performance, with respect to specific, widely used tests, are warranted. In keeping with this purpose, the present study specifically attempts to:

1. describe examinee acceptance of the *American College Test (ACT)* (American College Testing Program, 1960);
2. investigate the relationship of test acceptance to test performance.

Method

Subjects. At Kent State University, those entering freshmen who neglected to take the *ACT* on the national test dates are required to take the test on a residual basis during the summer orientation program. A random sample of 343 (206 men and 137 women) from the summer of 1963 residual testing program participated in this study. The most systematic difference between this group and those who took the *ACT* during the national testing program, is that these students made a late decision to enter college.

Questionnaire. After taking the *ACT* under the standard conditions, each *S* was asked to complete a questionnaire designed to assess certain examinee attitudes about the test. Three statements relating to each of the *ACT* subtests (English, mathematics, social studies, and natural science) were presented. Of the three statements, one referred to test length, one to test difficulty, and one to test quality. In anticipation of an overall questionnaire response set, half of the statements were given as "too long, too difficult, poor test" and half as "too short, too easy, good test." The same five response choices were available for each statement. These were: strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree.

ACT Scores. The *ACT* scores utilized in subsequent analyses were standard scores from each of the four subtests and the composite score, an average of subtest standard scores.

Results

The questionnaire data were first treated descriptively for each statement. Table 1 gives the frequency distribution in percentages, mean, and standard deviation of the responses for each statement. The responses were arbitrarily assigned numerical values of 1 through 5 in relation to degree of agreement with the statement, i.e., strongly agree = 1, strongly disagree = 5. The variable *n*'s are due to incomplete responses by some *Ss*.

To determine whether the specific questionnaire responses were

TABLE 1
Questionnaire Response Frequency Distribution, Means, and Sigmas for the Twelve Statements

Statement	Response Frequency (Percentages)					Mean	Sigma
	(1) Strongly Agree	(2) Agree	(3) Neutral	(4) Disagree	(5) Strongly Disagree		
1. The English test seemed too short.	1 ¹	11	41	44	4	3.40	.76
2. The English test is a good and fair test.	4	53	21	18	4	2.64	.95
3. On the whole, the English test seemed too easy.	0	6	40	49	4	3.51	.68
4. The math test seemed too long.	4	22	31	40	3	3.51	.93
5. The math test was too difficult.	5	27	32	32	4	3.03	.96
6. The math test is a poor test of skill and potential in mathematics.	2	15	35	44	3	3.29	.85
7. The social studies test is too short.	1	13	33	48	5	3.43	.81
8. The social studies test is a good test.	3	40	24	26	7	2.94	1.03
9. The social studies test is very hard.	3	22	41	32	2	3.07	.87
10. The natural science test is a poor test.	6	27	32	33	2	3.00	.96
11. The natural science test is too easy.	0	2	24	55	19	3.91	.73
12. The natural science test is too long.	3	25	38	33	1	3.02	.86

¹Rounded to the nearest percent.

independent of overall test performance, chi-square values were computed. For this purpose, the sample was trichotomized on the basis of the *ACT* composite scores. The low, middle, and high groups were defined by scores below 17, 17 to 20 inclusive, and above 20, respectively. Responses for each of the 12 statements were then tabulated by response category. Because of the presence of small theoretical frequencies in the extreme cells, the "strongly agree" and "agree" categories were combined, as were the two categories of disagreement. Thus, 3×3 contingency tables were organized for each statement. Table 2 reports the chi-square values obtained.

TABLE 2
*Chi-square Values for Test of Relationship of
Questionnaire Response to ACT Composite Score*

Statement	Chi-square	<i>p</i>
English . . . too short	5.19	n.s.
English . . . good and fair	2.60	n.s.
English . . . too easy	10.07	<.05
Math . . . too long	21.19	<.01
Math . . . too difficult	29.05	<.01
Math . . . poor test	4.50	n.s.
Social Studies . . . too short	17.12	<.01
Social Studies . . . good test	10.26	<.05
Social Studies . . . very hard	14.14	<.01
Natural Science . . . poor test	7.60	n.s.
Natural Science . . . too easy	2.57	n.s.
Natural Science . . . too long	10.77	<.05

d.f. = 4 in all cases.

Quintiserial correlation coefficients were computed for each questionnaire statement to determine the degree of relationship between subtest acceptance and performance on that subtest. The *ACT* subtest scores were organized into grouped frequency distributions with interval size three for easier computation. Table 3 gives the mean subscore, standard deviation, serial *r*, and the result of the "t" test under a hypothesized coefficient value of zero (Wert, Neidt, and Ahmann, 1954).

Discussion

An inspection of Table 1 in which one sees the responses of all the examinees to the 12 items of the questionnaire, fails in general to indicate a substantial criticism of the four subtests of the *ACT*. In

TABLE 3

Relationship between Response to Statement and the Related ACT Subtest Score

Statement	Subtest Mean	s.d.	r*	t	p
English . . . too short	18.07	4.72	-.19	2.98	<.01
English . . . good test			-.01	0.14	n.s.
English . . . too easy			-.21	3.53	<.01
Math . . . too long	17.58	5.78	.46	8.87	<.01
Math . . . too difficult			.54	10.89	<.01
Math . . . poor test			.17	2.88	<.01
Social Studies . . . too short	19.39	5.43	-.18	3.09	<.01
Social Studies . . . good test			.003	0.06	n.s.
Social Studies . . . very hard			.27	4.82	<.01
Natural Science . . . poor test	18.64	5.46	.03	0.49	n.s.
Natural Science . . . too easy			-.13	2.21	<.05
Natural Science . . . too long			.23	4.03	<.01

*Quintiserial correlation coefficient

regard to the English test, only 12 percent agree with the statement that the test seemed too short and only 6 percent that it seemed too easy. Seventy-eight percent are either in agreement or neutral with respect to the quality of the subtest. Twenty-six percent felt the mathematics test was too long; and 32 percent, that it was too difficult. Only 17 percent, however, felt that it was a poor measure of skill and potential in the field.

Fourteen percent felt that the social studies test was too short and 25 percent that it was too difficult. Two-thirds of the group were neutral or favorable in their disposition toward the quality of the subtest.

Thirty-three percent were critical of the quality of the natural science test and 28 percent agreed that it was too long. Only two percent felt that it was too easy.

When examinees were trichotomized with respect to their overall performance on the *ACT*, significant chi square values indicated that those who had achieved higher scores, in comparison with those who did less well, agreed that the English test was easy, did not view the mathematics test as too long or as too difficult, saw the social studies as too short but did not feel that it was either a very hard or a good test and did not view the natural science test as too long. Inspection of the cells of the chi square analysis suggests that the lowest scoring *ACT* group generally differed from those in the middle and high performance groups.

Rather similar findings were obtained when quintiserial correlations were computed between performance scores on each of the four tests and the agreement scale of the questionnaire items. Of the 12 items, 9 were significant at or beyond the .05 level. Only three items, those dealing with the quality (face validity) of the English, social studies and natural science subtests failed to yield relationships of reliable magnitude. In every case the examinees with higher scores in contrast to those with lower scores on each of the four subtests responded in such a way as to indicate that they felt less taxed by either the difficulty or number of subtest items.

Summary and Conclusion

Two hundred and six male and 137 female entering students at Kent State University were asked to complete a questionnaire concerned with the nature of the *American College Test* immediately after they had been examined on the instrument. Questionnaire items were concerned with the length, difficulty, and quality of each of four subtests. Questionnaire items were compared with overall performance on the *ACT* as well as upon performance on relevant subtests. These analyses clearly indicate covariation between test performance and perception of the test. In general, those students who score higher on the *ACT* are inclined to see the test as a task less demanding in length and in item difficulty than those who do less well.

REFERENCES

- The American College Testing Program. *Technical Report 1960-61*. Chicago: Science Research Associates, 1960.
- Anastasi, Anne. *Psychological Testing*. New York: Macmillan Company, 1954.
- Cronbach, L. J. *Essentials of Psychological Testing*. New York: Harper & Brothers, 1960. (a)
- Cronbach, L. J. "Is There Rest for the Test Weary?" *American Psychologist*, XV (1960), 665-666. (b)
- Cureton, E. E. "Validity." In E. F. Lindquist, (Editor), *Educational Measurement*. Washington: American Council on Education, 1951.
- Goslin, D. A. *The Search for Ability*. New York: Russell Sage Foundation, 1963.
- Mosier, C. I. "A Critical Examination of the Concepts of Face

Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 191-205.

Sarason, S. B. "Test Anxiety." *NEA Journal*, XLVIII (1959), 26-27.

Wert, J. E., Neidt, C. O., and Ahmann, J. S. *Statistical Methods in Educational and Psychological Research*. New York: Appleton-Century-Crofts, Inc., 1954.



PREDICTING GRADE POINT AVERAGE WITH THE *SRA TESTS OF EDUCATIONAL ABILITY*: A 13-MONTH FOLLOW-UP STUDY

WARREN S. BLUMENFELD¹

U. S. Naval Personnel Research Laboratory, San Diego, California²

Purpose

THE purpose of this investigation was to determine the longitudinal relationship between quotient scores obtained in the ninth grade on the *SRA Tests of Educational Ability, grades 6-9 (TEA)* (Science Research Associates, 1958) and a criterion of grade point average (GPA) at the conclusion of the tenth grade.

Predictor Variables

Four predictor variables were evaluated. These were the three sub-tests of the *TEA*—Language (*L*), Reasoning (*R*), and Quantitative (*Q*)—and the Total (*T*) score of the *TEA*. These indices were obtained in May of the ninth grade.

Sample

The participants in the study were 287 students—158 boys and 129 girls—who took the *TEA* in May of the ninth grade and subsequently completed the tenth grade at an urban high school in the San Francisco Bay area of California.

Criterion

The criterion used in this study was GPA in the tenth grade.

¹Data for this report were collected while the author was a member of the Test Department of Science Research Associates.

²Opinions expressed are those of the author and do not necessarily reflect those of the Navy Department.

GPA was obtained by calculating the arithmetic mean of grades of all courses taken in the two semesters of tenth grade. Grades were reported on an arbitrary scale where A = 5, B = 4, C = 3, D = 2, and F = 1. GPA was obtained in June of tenth grade; i.e., the time lapse between predictor collection and criterion collection was 13 months.

Results

Table 1 presents the means, standard deviations, and correlations between the five variables in the study.

TABLE 1
Means, Standard Deviations, and Intercorrelations of Variables^a
(*N* = 287)

Variable	M	SD	TEA L	TEA R	TEA Q	TEA T	GPA
TEA L	97.32	15.91	—	.67	.70	.92	.61
TEA R	97.34	14.75		—	.74	.87	.48
TEA Q	93.89	14.92			—	.88	.52
TEA T	95.87	15.07				—	.62
GPA	3.08	0.72					—

^aCorrelations of .18 are significant beyond the .01 level.

Conclusion

The data indicated that—at least for the particular sample studied—either *TEA T* or *TEA L* might be most successfully applied toward the prediction of the criterion of GPA over a school year's time.

REFERENCE

- Science Research Associates. *Manual of Directions for the SRA Tests of Educational Ability, Grades 6-9*. Chicago: Science Research Associates, 1958.

THE PREDICTION OF GRADES IN INTRODUCTORY PSYCHOLOGY FROM TESTS OF PRIMARY MENTAL ABILITIES¹

BISHWA NATH MUKHERJEE²

Indiana University

Background

THE predictive efficiency of various psychological tests in forecasting grades in introductory psychology has been examined by several investigators (Portenier, 1948; Jensen, 1950; Dysinger and Gregory, 1941; Bendig, 1958). However, excepting Ellison and Edgerton's (1941) study at the Ohio State University, no published report is available in the literature recording the validity of Thurstone's (1938a) original Primary Mental Abilities (PMA) Tests for prediction of grades in psychology at the freshman level.

Problem

This study was conducted to evaluate the degree to which semester grade in Introductory Psychology (P 101) as taught in the Southeastern Campus of Indiana University, Jeffersonville, Indiana, can be predicted on the basis of certain selected tests of PMA. The problem was to select from a battery of ten tests of PMA those variables which would yield the optimum estimate of the total grade points for the entire semester. These total points were the basis for final course grades.

¹Part of the financial support for this project came from the Graduate School, and computer time from the Research Computing Center of Indiana University. The valuable assistance of Miss Ruth Kobbe is gratefully acknowledged.

²Now at Patna University, Patna, India.

Predictor Variables

Nineteen tests were selected from the original 57 tests of *PMA* used by Thurstone (1938a) and published by him as a supplement (Thurstone, 1938b). The selected battery has been used previously by Vandenberg (1959) in his comparative study of the factorial structure of *PMA*. From these 19 tests, only eight served as the predictor variable for the present study. They are, namely, Vocabulary, Sound Grouping, Reading, Completion, Verbal Analogies, Verbal Classification, Reasoning, and Punched Holes. In addition to these tests, Thurstone and Thurstone's (1941) *Suffixes* and *Prefixes* as adapted recently by Educational Testing Service, Princeton, N. J. under the title *Word Endings* and *Word Beginnings*, respectively, were included in the battery.

Criterion Variables

In the first part of the study, the percentage marks received by each student on the first psychology quiz served as the criterion variable. In the third part of the study, the cumulative raw score received by a student for the entire semester was the criterion variable. Five regular examinations were given during the entire semester. Except for the final examination, each of the quizzes consisted of 100 points. The maximum possible score for the final two-hour examination was 150. Since Hilgard's *Introduction to Psychology*, (third edition) was the only text book for the P 101 course, all multiple choice items which appeared in these examinations were drawn from the Test-Item Booklets prepared by Hilgard.

Samples

In each of the three parts of the study, the students sampled were 87 freshmen enrolled in two different sections of P 101 class at the Southeastern Campus of Indiana University. For both the sections, the investigator himself was the teacher for the course. All testing for the study was done between November 13 and December 17, 1963. The mean age for the P 101 students of Fall 1963 semester was nearly 20 years. There were 51 male and 36 female students in the sample.

Procedure

1. Through use of IBM 709, product moment correlations were calculated among all possible pairings of variables described in Table 1. Then a stepwise multiple regression analysis (Efroymson, 1960) was made to select only those variables out of the predictor set composed of eight *PMA* tests which could yield the optimum estimate of the scores on the first psychology quiz. In the Efroymson stepwise procedure, which adds one variable to the prediction equation at a time, the beta weight for each variable is tested against zero by the *F* test of variance at a given significance level. If the beta weight for the variable is statistically not significant, then the variable is dropped from the multiple regression equation. The analysis was performed on an IBM 709 Computer located at Indiana University, Bloomington; the program was BIMD 09, obtained from the UCLA Medical Center.

2. Through use of the regression equation which turned out to be the most optimal, the predicted score for each student was calculated. The estimated scores, however, were compared not only to the scores actually obtained by the students in the first quiz but also to their total semester point percentages. This was done in order to see how far tests which can successfully predict the scores on first quiz can efficiently predict the entire semester points. This step formed the second part of the study, the main purpose of which was to determine how reliable was the regression equation obtained in Part I. The standard error of estimate, i.e., the standard deviation of the discrepancy scores between the actual total scores converted in percentage and the predicted scores in percentage, was computed which gave an index as to the predictive efficiency of the selected variables.

3. In the final part of the study, a step-wise regression was made, in which all the psychological tests including the *Word Endings* and *Word Beginnings* as the predictors and the total psychology scores as the criterion were employed. The program used for this purpose was Spiegel and Bock's (1963) multiple step-wise regression routine prepared for LGP-30. The order in which the variables entered the regression was determined *a priori* in terms of their correlation with the total psychology scores. The program computes

the regression equation, the coefficient of determination, and the step-down F statistics at each step of analysis.

Results

The intercorrelations among all the tests and the five psychology quizzes appear in Table 1. In Table 2 are reported the multiple regression weights for the selected variables which contributed significantly in predicting scores on the first psychology quiz. Table 3 shows the main results of step-wise regression analysis.

Discussion

The results of this study very strongly suggest that at least two tests of *PMA* can be used to predict academic performance in psychology at the introductory level. The tests are, namely, vocabulary and sound grouping. The Vocabulary Test turned out to be the best single predictor of all the tests of *PMA*. As is well known, vocabulary plays an important role in most college courses. It is no surprise therefore to find that students having a better vocabulary scored higher in the psychology examinations. The Sound Grouping Test has been found to be a good measure of perceptual (auditory) speed. Its predictive value in forecasting psychology grades does not seem to be very high. Nevertheless, for the sample tested, it can improve the prediction quite appreciably.

The study also demonstrates that the regression equation obtained for the prediction of first psychology quiz score worked equally well for the prediction of total scores in psychology. As a matter of fact, the standard error of estimate for the latter case was appreciably lower. When the actual total points in psychology converted in terms of percentage were compared with the corresponding predicted scores computed from the regression equation obtained in Part I, the mean discrepancy was 1.86. The standard deviation of these discrepancy scores, i.e., standard error of prediction, was 7.47 compared to the standard error of estimate of 10.54 obtained for the prediction of first psychology quiz score.

For the sample studied, it seems that the following regression equation where only two variables enter will give an optimum estimate of the total psychology scores at the introductory level:

$$Y_t = .507 X_{1t} + .268 X_{2t}$$

TABLE 1

Product Moment Intercorrelation among Selected Tests of PMA, Word Beginnings, Word Endings, and Scores on Different Psychology Quizzes*

Variable	Psychology Quiz										
	Voca	SG	Read	Comp	VA	VC	Reas	PH	WB	WE	Total
Vocabulary	1.00	.46	.60	.54	.47	.30	.25	.28	.30	.33	.63
Sound Grouping		1.00	.47	.26	.40	.48	.21	.19	.30	.55	.50
Reading			1.00	.44	.41	.26	.33	.26	.22	.33	.48
Completion				1.00	.44	.26	.27	.33	.25	.26	.42
Verbal Analogies					1.00	.31	.46	.28	.31	.34	.38
Verbal Classification						1.00	.10	.21	.05	.13	.32
Reasoning							1.00	.33	.18	.20	.30
Punched Holes								1.00	.29	.34	.21
Word Beginnings									1.00	.55	.28
Word Endings										1.00	.42
First Quiz										.44	.75
Second Quiz										1.00	.73
Third Quiz										.63	.77
Fourth Quiz										1.00	.68
Fifth Quiz											1.00

* $r = .28$ required for significance at the .01 level.

TABLE 2

*Results of Various Stepwise Regression Analysis for
Predicting Scores on the First Psychology Quiz*

Variables retained for the analysis	Predictor Variable omitted*	Criterion F -level	Variables Selected	Beta Coeff.	St. Err. of Coeff.	Multi R	St. Err. of Estim.
All PMA tests	WB, WE, & Psy Scores	4.5	Vocabulary	.623	.099	.560**	10.66
All PMA tests	WB, WE, & Psy Scores	3.4	Vocabulary	.536	.111	.601**	10.54
			Sound Grpg	.174	.103		
All PMA tests,	Psychology	3.4	Vocabulary	.458	.113		
WB & WE	Quizzes		Word Endings	.361	.108	.672**	9.49
			Sound Grpg	.094	.103		

*Abbreviations explained in text.

** $p < .01$.

where Y_i is the total psychology score for individual i and X_{1i} and X_{2i} denote respectively the scores on the Vocabulary and Sound Grouping Tests of Primary Mental Abilities.

The data suggest that it would be an unwise use of time to administer the entire PMA battery for the purpose of predicting suc-

TABLE 3

Regression Coefficients, Multiple Correlation between Predictor Sets, and the Total Psychology Scores at Different Steps of Analysis Together with the F Statistics

Step No.	Predictor Variables	Regression Coefficient	Multiple R	F -level
1	Vocabulary	.5067		
	Sound Grouping	.2683	.6745	8.718**
2	Vocabulary	.5364		
	Sound Grouping	.1037	.7110	.7110
	Reading	.2710		
3	Vocabulary	.5153		
	Sound Grouping	.1024	.7175	7.681*
	Reading	.2544		
	Word Endings	.1047		
4	Vocabulary	.5065		
	Sound Grouping	.0434	.7235	.022
	Reading	.2757		
	Word Endings	.1045		
	Completion	.1072		

* $p < .05$.

** $p < .01$.

cess in beginning college psychology. Even the inclusion of Reading Test appears prohibitive. For the sample studied inclusion of other PMA tests might increase the multiple correlation a little bit but in view of the fact that multiple regression weights become less useful for a new sample as the number of predictors increases, it seems desirable to use the Vocabulary and the Sound Grouping Tests only.

Of course, crossvalidation studies are needed to ascertain whether the results reported here would still hold true for new samples. Such studies are being planned for the future.

Summary

Using stepwise multiple regression technique, it was found that only the Vocabulary and the Sound Grouping Tests of PMA could predict satisfactorily the total grade points in beginning college psychology. A test of vocabulary turned out to be the best single predictor.

REFERENCES

- Bendig, A. W. "Comparative Validity of Empirical Temperament Test Keys in Predicting Student Achievement in Psychology." *Journal of Educational Research*, LI (1958), 341-348.
- Dysinger, D. W. and Gregory, W. S. "A Preliminary Study of Some Factors Related to Student Achievement and Grades in the Beginning Course in Psychology." *Journal of General Psychology*, XXIV (1941), 195-209.
- Efroymson, M. A. "Multiple Regression Analysis." In A. Ralston and H. S. Wilf (Eds.), *Mathematical Methods for Digital Computers*. New York: Wiley, 1960.
- Ellison, M. L. and Edgerton, H. A. "The Thurstone Primary Mental Abilities Tests and College Marks." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, I (1941), 399-406.
- Jensen, B. T. "Evaluating Achievement in Psychology." *American Psychologist*, V (1950), 343. (Abstract.)
- Portenier, L. G. "Predicting Success in Introductory Psychology." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII (1948), 117-126.
- Spiegel, D. K. and Bock, R. D. "Multivariate Step-wise Regression Analysis: A Program for the General Precision LGP-30." Research Memorandum, University of North Carolina Psychometric Laboratory, Chapel Hill, N. C., July, 1963.
- Thurstone, L. L. "Primary Mental Abilities." *Psychometric Monographs*, No. 1, 1938. (a)
- Thurstone, L. L. *Primary Mental Abilities* (Supplement). Chicago: University of Chicago Press, 1938. (b)

- Thurstone, L. L. and Thurstone, T. G. "Factorial Studies of Intelligence." *Psychometric Monographs*, No. 2, 1941.
- Vandenberg, S. G. "The Primary Mental Abilities of Chinese Students: A Comparative Study of the Stability of a Factor Structure." *Annals of the New York Academy of Sciences*, LXXXIX (1959), 257-304.

INTELLECTIVE PREDICTORS OF SUCCESS IN NURSING SCHOOL¹

JON M. PLAPP, GEORGE PSATHAS

Washington University

AND

DANIEL V. CAPUTO

Washington University School of Medicine

RATE of attrition from schools and colleges continues to present a challenge to selectors and educators. Although aptitude tests and other predictive measures are used in most initial selection procedures, large numbers of dropouts still occur during nursing training. Most take place in the first year (Tate, 1961). Continued investigation of the effectiveness of predictive instruments and of the reasons for attrition is necessary. In a review of the literature on dropouts from colleges, Summerskill (1962) considered biological and social, motivational, illness-and-injury, adjustment, and academic-performance correlates of college success. Of these, academic factors (secondary school performance, scholastic aptitude test scores, and academic performance in college), proved to be the most effective predictors of dropouts. Past academic performance and achievement and aptitude test scores have been studied most frequently; the other factors, though deserving of more intensive study, have been both less readily obtainable and less quantitative. These factors pose complex measurement and validity problems in them-

¹ This paper is a partial report of a research project "Role Differentials and Nursing Ideology," John A. Stern and Albert F. Wessen, Co-Principal Investigators; the research has been supported by PHS Grant NU-00050. A portion of the computation was done with support provided by the Washington University Computer Center under National Science Foundation Grant G-22296. The authors wish to express their appreciation to Joanne Schwartz for her valuable assistance.

selves. Academic and test measures appear to be the most consistently accurate predictors of dropouts.

Studies of the relation between intelligence and scholastic achievement in elementary and high schools indicate that there is a reliable correlation between these variables. In college and other highly selected samples, however, this relationship would be expected to be attenuated, with other than intellectual factors assuming more importance and taking up a greater proportion of the total variance associated with success.

Among the research publications reporting reliable and significant correlations between intellectual predictors and academic performance, there has been a considerable number specifically concerned with the prediction of success in nursing school. A review of research in this area by Taylor *et al.* (1963) indicated that among the most frequently studied and apparently effective correlates of certain criteria of nursing school performance were IQ tests, aptitude tests, and rank in high school class. It was also reported by Jacobs (1959) that scores on pre-admission tests and high school rank were used most often by the nursing schools as criteria of selection. The most frequently studied criteria of success in nursing school have been measures of academic performance and of continuance in school. Clinical grades have been studied much less frequently than academic or theory grades. The general finding has been that correlations between intellectual predictors and criteria of clinical performance are lower than are correlations between intellectual predictors and academic performance (Taylor *et al.*, 1963).

One of the aims of the present study was to compare the efficacy of the intellectual predictors selected—high school rank (*HSR*), the 1937, *Gamma AM Form* of the *Otis Quick-Scoring Mental Ability Tests* (*OTIS*), and the *Scholastic Aptitude Test* of the College Entrance Examination (*SAT*)—alone and in combination, in accounting for the performance of nursing students on the criteria of continuance, academic performance, and clinical performance. The comparison was also expected to reveal which one of the three generic measures of ability (past performance, general intelligence, and aptitude tests) shows the greatest relationship to each of the three criteria. The highest correlations should obtain between those variables which seem to have most in common—i.e., between a measure of past academic performance (such as *HSR*)

and academic grades in nursing school. An additional reason for this expectation is the higher reliability the *HSR* measure is assumed to have, since it represents four years of high school performance, as opposed to the relatively short times covered by performance on the *SAT* and the *OTIS*. Because of the variations in size of school attended, *HSR* may be limited in its usefulness as a predictor. In order that this effect might be mitigated, a self-rating (*SR*) made by each student of her past high school performance in terms of a four point scale of "poor," "average," "good," or "excellent" was also used. Such a rating should not be so limited by size of class as should actual *HSR*; hence the correlations between *SR* and the criteria would be expected to be higher than those between *HSR* and criteria.

Some attempt was also made in the present investigation to deal with the problem of the relevance of the criteria studied to actual clinical-nursing performance. It has been suggested (Taylor *et al.*, 1959) that the criteria of adequacy of performance for a student nurse may bear little relationship to the criteria of adequacy of Registered Nurse (R. N.) performance. This apparent lack of relationship raises the possibility that the successful nursing school applicant is more likely to be selected because she is expected to perform successfully in nursing school than because of her probable capability for fulfilling the actual role of a practicing R. N. This problem is a general one in applied fields, since often selection for entrance into training rests solely on measures of intellectual performance with little concern for the individual's abilities in the applied area. Accordingly, in the present study, correlations were obtained between the predictors and the criterion of clinical nursing grades, which were assigned in the final quarter of the freshman year. Clinical performance, like that required of the R. N., is more directly involved in these grades than in grades from academic courses. It was predicted that correlations between the predictors and clinical grades would be lower than those correlations between the predictors and academic grades obtained at the same time.

Another factor which has received little attention to date was the question of whether the predictive power of the intellectual variables would persist over time. The assumption of predictive stability, based on the early research on intelligence tests, has persisted, though it has not specifically been investigated in normal adoles-

cents or adults. The aim here was to discover whether the proportion of the variance of grades in first-quarter academic nursing courses found to be attributable to the intellectual predictors would continue to show a stable relationship over a period of approximately one year or whether the predictive efficacy of the three intellectual measures would be affected by time. If the predictors are successful only in accounting for variance in the first but not the final quarter grades, the possibility that immediate scholastic abilities rather than "true" intellectual potential was the basis for the selection of students becomes a more plausible interpretation.

Examination of the criterion of continuance-discontinuance suggests that it too may not be a homogeneous criterion. The relationship between the intellectual predictors and this criterion would be expected to be higher for the first year of school than for the second. The majority of "casualties" arising from poor academic performance would be expected in the first year; in later years, after this selection had occurred, other facts such as motivation, marriage, and personality deficits would assume greater importance as reasons for discontinuance. The question to be considered here, then, was whether first and second year "dropouts" represented two distinct groups.

The hypotheses of the present study are the following.

1. Intellectual predictors and academic course grades will show a significant positive correlation.
2. Intellectual predictors will be significantly correlated with the criterion of continuance-discontinuance.
3. Correlations between the intellectual predictors and continuance-discontinuance during the first year of nursing school will be higher than will be the correlations between these predictors and continuance-discontinuance during the second year.
4. Correlations between intellectual predictors and academic grades will be higher than will be the correlations between these same predictors and the continuance-discontinuance criterion.
5. Time will have an attenuating effect on correlations between intellectual predictors and scholastic performance. Correlations between the predictors and academic course grades in the first quarter of nursing school will be higher than will be correlations between the predictors and fourth quarter academic course grades for the same group of subjects.

6. Correlations between the intellectual predictors, singly or in combination, and the criterion of academic course grades will be higher than will be correlations between the same predictors and clinical grades, when both sets of grades are obtained at the same point in time and for the same group of subjects.

7. A predictor consisting of a measure of past scholastic performance, *HSR*, will correlate more highly with first and fourth quarter academic course grades than will other, more general intellectual predictors such as the *OTIS* and *SAT*.

a. The *SR* will correlate more highly with scholastic performance than will *HSR*.

8. The predictors used in the present investigation will be positively intercorrelated.

Method

In general, the methodology involved obtaining intellectual measures of the subject group immediately prior to entrance into the nursing school and obtaining criterion measures during the subsequent two years.

Subjects

The subjects of the present study were the 79 students who comprised the 1962 freshman class at the nursing school of a general hospital in St. Louis. The school offered a three-year diploma program leading to eligibility to take the R.N. licensing examination. The mean age of these girls was 18 years. Most had entered the nursing school directly following graduation from one of the high schools in the greater St. Louis area. Table 1 identifies the sample

TABLE 1
Scores of Subjects on Predictor Variables

Predictor*	Mean	S.D.	Self Rating*	N	%
OTIS (IQ)	113.90	5.81	"Poor"	0	0
SAT (Composite Score)	295.31	43.08	"Average"	22	28.57
HSR (Percentile in Graduating Class)	70.59	18.43	"Good"	48	62.34
			"Excellent"	7	9.09

*N is 77 for each predictor.

in terms of the variables of present interest, *Otis IQ* tests, Scholastic Aptitude Test, High School Rank, and Self Rating of high school performance.

The intellectual predictors employed were the scores of the students on the *Otis IQ* test (*OTIS*), the Scholastic Aptitude Test (*SAT*); and the rank of each student in her high school class (*HSR*). The latter was a rank of the student's performance during her four years of high school formed by ranking each of the students' high school ranks within the group of nursing students. The fourth predictor was a subjective self rating (*SR*), made by each student of her high school record, using the categories: "poor," "average," "good," or "excellent."

These measures had all been obtained prior to the beginning of the students' freshman year.

The criterion variables to which the above predictors were to be related consisted of the following: information as to the dropout status of each student obtained on two occasions, at the conclusion of the first year, and at the conclusion of the second year; the grades obtained at the end of the first quarter in academic courses (fundamentals of nursing, chemistry, anatomy, and sociology); the grades in the sole academic course taken during the fourth quarter (normal nutrition); and the fourth quarter grades in clinical (medical-surgical) nursing. In order for an assessment to be made of the efficacy of the predictors in accounting for the variance in first quarter as compared with fourth quarter academic grades, correlations between the predictors and these criteria were computed only for that group of students who obtained grades on both occasions, i.e., for those students who remained in school through the fourth quarter final examinations. Similarly, because clinical nursing grades were given only after the fourth quarter had been completed, correlations with this criterion had to be restricted to those students who remained in school at the end of the fourth quarter. However, correlations were also computed between the predictors and the criterion of first quarter grades for all students who took the first quarter examinations (i.e., including those who later dropped out).

The number of students who had left school by the end of the second year was 27, with 50 remaining in school. Of these 27 dropouts, 17 had occurred during the first year and ten during the second. The reduction of the total sample from 79 to 77 was the

result of incomplete information on the predictor variables for two subjects.

Point-biserial correlations were computed between each predictor and each criterion, with the exception of the correlations involving the self-rating predictor. In this latter instance phi-coefficients were obtained. For the point-biserial correlations, the predictors comprised the continuous, and the criteria the discrete, variables. Grades were divided above and below the median to form discrete variables. Through use of the Pearson r the three objective predictor variables were also intercorrelated.

Results and Discussion

In general, the hypotheses of the present study were confirmed. As can be seen from Table 2, significant correlations were obtained between each of the predictors and at least one of the criteria of nursing school performance. The only two criteria to which none of the predictors bore a significant relationship were those of dropout status during the second year of nursing school, and fourth quarter academic grades. The first two hypotheses of the present study, which were concerned with predicting significant correlations between intellectual measures and criteria of course grades and continuance-discontinuance, were thus supported, with these exceptions. As expected, it seems likely that after the first year complex motivational and personal factors come to replace intellectual factors as the major reasons for dropouts from school. The efficacy of the intellectual variables in predicting first year dropouts thus seems to be reliably established; their usefulness in predicting second year dropouts, however, appears negligible.

This finding raises the possibility that significant correlations between intellectual factors and dropout status in nursing school are due to the covariation of predictors and criteria during the first year only. General statements implying an enduring and stable covariation throughout the three or four years of nursing school are called into question by the present findings. The need for developing predictors which will account for continuance and discontinuance after the first year is also highlighted.

As predicted in hypothesis three, correlations between predictors and grades were generally higher than between predictors and the criterion of continuance-discontinuance. This outcome was most

TABLE 2
Correlations between Intellectual Predictors and Criteria of Nursing School Performance^a

Predictors	Criteria						
	All: "Outs" vs. "Ins" (27) (50)	1st Yr: "Outs" vs. "Ins" (17) (50)	2nd Yr: "Outs" vs. "Ins" (10) (50)	Grades 1st Quarter Academic (N = 77)	Grades 1st Quarter Academic (N = 60)	Grades 4th Quarter Academic (N = 60)	Grades 4th Quarter Clinical (N = 60)
OTIS	.11	.20	.06	.30**	.36**	.01	.22
SAT	.23	.26*	.11	.27*	.25	.12	.28*
HSR	.22	.30*	.04	.44**	.36**	.14	.09
OTIS & SAT	.19	.25*	.03	.32**	.33**	.07	.27*
OTIS & HSR	.21	.32*	.01	.46**	.45**	.10	.19
SAT & HSR	.29*	.35*	.10	.45**	.39**	.17	.24
OTIS, SAT, & HSR	.25*	.33*	.04	.44**	.42**	.12	.25
SR	.42**	.44**	.22	.46**	.20	.22	.22

*Significant at .05 level.

**Significant at .01 level.

^aAll correlations are point-biserial coefficients, with the exception of those between SR and criteria, which are phi coefficients.

clearly the case for correlations between predictors and grades in first quarter courses (which were largely academic or scholastic in character) as compared with dropout-persistence status, for the two years combined. Correlations between the intellectual predictors and the first quarter grades were significant at the .01 level with the single exception of the *SAT* correlation which attained the .05 level of significance. By contrast, the correlations between predictors and dropout-continuance status (with all dropouts included) reached the .01 significance level only for *SR* and the .05 level for the *SAT* and *HSR* and for the *OTIS*, *SAT*, and *HSR* combinations. A greater number of significant correlations and generally higher levels of significance were thus obtained between the intellectual predictors and first quarter grades than between these predictors and continuance-discontinuance (over a two-year period).

The fifth hypothesis regarding the influence of time on the correlations, was also supported. As was the case with the prediction of dropout-persistence status, there was a marked attenuation of the levels of relationship between predictors and grades with the passage of time. As stated above, for the total group of 77 subjects, significant correlations were obtained between each predictor and the criterion of first quarter grades. Only two of these significance levels, involving *SAT* and *SR* predictors, were reduced when the group was restricted to only those 60 subjects who had obtained first quarter grades and also had completed the fourth quarter of the freshman year. Mere reduction of the range of talent, then, had little effect on the majority of the correlations. On the other hand, when these correlations with first quarter academic course grades are compared with the correlations with the academic course grade obtained at the end of the fourth quarter, marked reductions in all but one correlation are observed. Such attenuation might be based on the fact that while four academic courses were taken by students during the first quarter, only one was taken during the fourth quarter. The fourth quarter grade is, therefore, a less reliable measure. The only relationship which remained stable was that between *SR* and the criteria; a possible explanation of this stability is the fact that phi, the measure of relationship employed here, is a relatively crude index, less sensitive to actual variations in levels of relationship than the point-biserial correlations used to assess the relationships between the seven remaining predictors and the cri-

teria. Inquiry into the nature of the academic courses used as criteria of first and fourth quarter performance revealed much commonality: all were theoretical, made use of text books, required note taking and study, and included final examinations. Apparently the reason for the marked attenuation of correlations over time could not have been the result of obvious differences between the academic courses taken in first and fourth quarters. This finding is in contra-distinction to the assumptions underlying the use of intellectual predictors in general, for their usefulness depends on their capabilities for predicting performance not only in the first year, but also throughout the training period and after. Aside from the possibly lower reliability of the fourth quarter grade, it would seem that certain changes in the motivations of the students, or in the ways in which faculty members evaluated students, may have produced the reduction in size of coefficients. A greater stress on clinical performance with length of time in school would help to explain not only the attenuation of predictive coefficients of correlation with academic performance, but also the contrasting higher level of relationship which was found between certain predictors and clinical performance at the conclusion of the fourth quarter of nursing school. For, contrary to hypothesis six, correlations, for the same group of subjects, between the predictors and clinical grades were generally higher than were the correlations between predictors and academic course grades also obtained in the fourth quarter.

The hypothesis that the *HSR* in comparison with other measures would be a relatively more effective predictor of scholastic performance, and more particularly, of first quarter academic grades (hypothesis seven) was not clearly supported. As can be seen from Table 1, the *HSR* correlation with the criterion was .44, whereas the others were: *OTIS*, .30; *SAT*, .27. However, since evaluation, by means of *t*-tests, of the significance of the differences between these correlations, revealed that none was significant, the hypothesis can only be said to be supported by a trend in the data. Nor did the supposedly less restricted measure, *SR*, correlate with the criterion to a significantly greater degree than did *HSR*. Hence, this corollary hypothesis (hypothesis 7a) also failed to obtain confirmation.

Examination of the correlations between the predictors and clinical grades reveals that *SAT* was the single predictor most strongly related to this criterion, whereas *HSR* was the weakest predictor.

These relationships are in direct contrast to the correlations between these two predictors and the first quarter academic grades of the same group of 60 subjects. In the latter instance, correlations involving *HSR* were uniformly higher than were those involving *SAT*; in fact, the relationship between *SAT* itself and academic grades failed to attain significance. Thus, although *HSR* was one of the most successful predictors of academic course grades in the first quarter of nursing school, it had ceased to be an effective predictor by the time fourth quarter grades were obtained. By contrast, although the *SAT* was a relatively unsuccessful predictor of first quarter grades, and as ineffective in predicting fourth quarter academic course grades as any other predictors, it was able to predict significantly fourth quarter clinical course grades. The two predictors appear to be measuring factors which may be considerably divergent. It is of interest that a test requiring two hours and twenty minutes of performance is a more successful predictor of clinical grades than is a measure representing four years of high school performance.

The final hypothesis, regarding the existence of positive interrelationships among the four predictors used in the present study, was clearly supported. Table 3 presents intercorrelations of the predictors and indicates that all reached significance. In accordance with expectations, the correlation between *HSR* and *SR*, being based on interdependent data, was high, as was the correlation between the *OTIS* and *SAT* measures. The degree of significance attained between the *OTIS* and the other predictors indicates the value of this relatively short and easily administered test.

TABLE 3
*Intercorrelations between Predictors**

	<i>SAT</i>	<i>HSR</i>	<i>SR</i>
<i>OTIS</i>	.65**	.28*	.33**
<i>SAT</i>		.26*	.23*
<i>HSR</i>			.57**

*Significant at .05 level.

**Significant at .01 level.

*All correlations are Pearson product-moment coefficients, with the exception of those involving *SR*, which are point-biserial coefficients.

No specific predictions were made concerning relationships between combinations of the intellectual variables and the criteria. Examination of these relationships, and comparison with the levels

of correlation between individual predictors and the criteria, reveals that there was no significant advantage to using combined predictors, and that no individual predictors had a *general* superiority over any other. All predictors shared a certain amount of common variance, as can be seen in Table 3, moreover, either nearly all or almost none of the predictors were significantly related to any particular criterion. Correlations between predictors and grades in the clinical nursing course were the only clear exceptions to this statement: in predicting this criterion correlations ranged from .09 for *HSR* to .28 for *SAT*. In general, then, it must be concluded that although there was no advantage in employing combined predictors, the use of a number of different single predictors did have some merit.

Conclusions

The broader implications of the present findings are related to certain assumptions made by those who employ batteries of intellectual tests and records of past scholastic performance in selecting applicants for advanced training or for occupations. The basic assumption is of course that the use of these procedures will make for more effective selection. Effectiveness of selection is judged initially in terms of how adequately successful applicants respond to the new situations encountered following admission—for example, how well students perform in course work or carry out designated assignments. Those students who fail to master the expected material, and those students who, for whatever reason, terminate prematurely their associations with the training or employing institution, can be considered to represent examples of failures in selection. A second implicitly held assumption of selectors is that the predictive power of the tests used will remain consistent over time. Closely related to this is the further assumption that the selection battery will predict equally well success in the initial and later stages of the training program. Usually the choice of information to be used in the selection program is determined purely by the demonstrated validity of the tests in predicting performance in the initial stages of training. The findings of the present investigation cast doubt on the appropriateness of this practice, for they demonstrate that the assumptions regarding consistency in predictive power of the selection battery are not necessarily valid. The need is great for further examination of the validity of selection batteries in predicting suc-

cess during later stages of training and in actual occupational performance, as well as for the development of tests which will perform these functions. With such research will also come more adequate descriptions of occupational role performance, since only after the criteria have been adequately described will effective predictors be discovered.

Summary

The efficacy with which four intellectual predictors, singly and in combination, predicted success in nursing school, was assessed. The predictors consisted of an intelligence test, a scholastic aptitude test score, high school rank, and self-rating of high school performance. Criteria of success were the students' grades in scholastic and clinical courses, and their status as "dropouts" or "stayins." Data were collected on 77 subjects.

Generally, significant correlations were found between predictors and criteria. When these relationships were examined in detail, however, it was found that significant correlations existed between the predictors and dropout-continuance status in the first year of nursing school but not in the second year, and that the predictors correlated significantly with first quarter, but not with fourth quarter grades in academic courses. The Scholastic Aptitude Test was the only single predictor to correlate significantly with fourth quarter grades in the clinical nursing course.

These findings served to question the following two assumptions commonly made by the users of intellectual test batteries: (a) the predictive power of the tests remains consistent over time, and (b) prediction of academic performance can be equated with prediction of clinical performance.

REFERENCES

- Jacobs, J. H. "The Nursing School Applicant." *Careers in Nursing Committee, Special Report Series No. 5*. Philadelphia, Pa.: Southeastern Pennsylvania League for Nursing, 1959.
- Summerskill, J. "Dropouts from College." In N. Sanford, *The American College*. New York: John Wiley & Sons, Inc., 1962.
- Tate, Barbara. *Study of Attrition Rates in Schools of Nursing*. New York: National League for Nursing, Inc., 1961.
- Taylor, C. W., Nahm, Helen, Loy, Lorraine, Harms, Mary, Berthold, Jeanne, and Wolfer, J. A. *Selection and Recruitment of Nurses and Nursing Students*. Salt Lake City: University of Utah Press, 1963.



THE PREDICTIVE VALIDITY OF A BATTERY OF DIVERSIFIED MEASURES RELATIVE TO SUCCESS IN STUDENT NURSING

WILLIAM B. MICHAEL

University of California, Santa Barbara

RUSSELL HANEY

Los Angeles, California

AND

STEPHEN W. BROWN

University of Southern California

Problem

For a sample of 118 freshman trainees in a student nursing program at the Los Angeles County Hospital during the 1963-1964 academic year, it was the primary objective of this investigation to determine indices of predictive validity for both cognitive and non-cognitive variables relative to 12 criterion measures. The secondary purpose was to obtain what amounted to crossvalidation data for several of the predictor variables that had been employed in two previous studies by Haney, Michael, and Gershon, (1962) and by Michael, Haney, and Gershon (1963) as well as in other previous investigations reported in the bibliographies of these two studies. Most of the predictor variables cited have been described in these two previous articles—especially in the earlier paper. The criterion variables, many of which have also been described in these same two references, are self-explanatory in terms of their titles in Table 1.

The only additional information that seems necessary to provide in the description of the predictor variables is that (1) the *Survey of Space Relations Ability (SSRA)* is a test which was prepared by Floyd Ruch and Harry W. Case and published (in 1944) by the

California Test Bureau, (2) only those *MMPI* scales were included in Table 1 for which two or more predictive validities coefficients significant at or beyond the .05 level were obtained, and (3) the Ward Adjustment Rating Scale consisted of 12 instead of five items as reported in the study by Michael, Haney, and Gershon (1963)—an instrument for which scale values on the items were summed to furnish a total score.

Statistical Treatment

Through use of an IBM 7090 correlation program at the Western Data Processing Center at UCLA, product-moment coefficients of correlation were computed among all possible pairings of variables cited in Table 1. Scores had been converted to stanines except in the instance of variables 25 and 26, for which a five-step scale was employed. Despite the existence of experimental dependence among the various scales of the *MMPI*, three factor analyses involving use of Kaiser's (1959) varimax method of rotation were undertaken for (1) the entire matrix of variables with the exclusion of total scores on the *California Reading Test* and the *California Mathematics Test*, (2) the matrix of all predictor variables (including all *MMPI* scales), and (3) the set of all criterion variables including each of the items on the Ward Adjustment Scale. Although the results of the factor analyses will not be described in detail, a few highlights will be briefly mentioned. It should be emphasized that the predictive validities are probably minimal in value, since there was a substantial restriction in range in the 118 scores when their spread was compared with that of the applicant group of nearly 500 students. In addition to a lack of presentation of means and standard deviations of each of the variables in Table 1, no corrections for restriction of range are reported.

Results

The following major findings may be summarized from the entries in Table 1.

1. Among the cognitive tests used, the comprehension score on the *California Reading Test* was the most valid predictor of grades in two nursing oriented courses, in two psychology courses, and in the nutrition course. It was also a significantly valid predictor of success in physiology and microbiology. The portion of the *California*

TABLE 1

Predictive Validity Coefficients of Measures for the 1963-1964 Nursing Class at the Los Angeles County Hospital (N = 118)

Predictors (Variables 1-15) and Criterion Measures (Variables 16-27)	Validity Coefficients of Predictors along with Intercorrelations of Criterion Measures*											
	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)
(1) Calif. Reading Test (Total)	29	19	10	15	03	29	24	19	13	25	17	15
(2) Calif. Reading Test (Vocabulary)	26	19	12	14	09	31	22	19	15	30	19	16
(3) Calif. Reading Test (Comprehension)	42	33	19	20	16	38	36	12	32	26	24	15
(4) Calif. Math Test (Total)	17	12	16	16	37	07	10	09	12	14	11	02
(5) Calif. Math Test (Reasoning)	16	07	12	06	27	11	05	08	09	08	12	-07
(6) Calif. Math Test (Fundamentals)	26	22	21	25	47	16	19	17	20	22	16	13
(7) EAST No. 5 Space Visualization	-07	19	00	-04	03	-20	-15	-04	-16	-04	-15	-07
(8) Survey of Space Relations Ability (SSRA)	-01	-03	-07	-09	00	-09	-11	10	-04	05	-04	-01
(9) High School Chemistry (Two Semester GPA)	28	18	21	14	25	17	24	26	32	22	31	03
(10) High School Solids (GPA)	39	48	22	22	39	36	33	37	39	44	51	22
(11) High School GPA	33	34	25	20	40	28	18	31	25	34	37	33
(12) MMPI-F Scale	-19	-17	-22	-05	-25	-25	-23	-18	-28	-13	-10	-13
(13) MMPI-Pd + .4K Scale	-08	-20	-07	03	-26	-12	21	-15	-14	-16	-31	-09
(14) MMPI-Sc + 1K Scale	-10	-23	-14	-07	-20	-20	-10	-16	-19	-22	-28	-09
(15) MMPI-Ma + .2K Scale	-32	-25	-23	05	-13	-29	-28	-19	-32	-21	-20	-08
(16) Nursing I Grades	—	74	35	24	45	64	60	55	65	59	54	42
(17) Nursing II Grades	74	—	30	36	44	68	70	62	71	70	68	32
(18) Orientation to Nursing Grades	35	30	—	41	31	37	36	24	34	30	23	25
(19) History of Nursing Grades	24	36	41	—	22	22	27	14	26	41	33	18
(20) Arithmetic in Nursing Grades	45	44	31	22	—	36	32	39	41	46	40	22
(21) Psychology I Grades	64	68	37	22	36	—	73	52	68	56	47	26
(22) Psychology II Grades	60	70	36	27	32	73	—	60	64	63	57	22
(23) Anatomy Grades	55	62	24	14	39	52	60	—	56	63	50	30
(24) Nutrition Grades	65	71	34	26	41	68	64	56	—	65	60	29
(25) Physiology Grades	59	70	30	41	46	56	63	63	65	—	69	34
(26) Microbiology Grades	54	68	23	33	40	47	57	50	60	69	—	22
(27) Ward Adjustment Rating Scale	42	32	25	18	22	26	22	30	29	34	22	—

*Uncorrected for restriction of range coefficients of .18 and .24 are significant at the .05 and .01 level.

Mathematics Test concerned with fundamentals was the most valid indicator of performance in a course of Arithmetic for Nurses (as might be expected), as well as a significantly valid predictor of standing in four nursing oriented courses (variables 16-19), in nutrition, and in psychology. The other test measures in the cognitive domain added little to the prediction of the criterion variables—especially in the instance of the two tests concerned with spatial abilities.

2. Although all three measures representing high school achievement were promising predictors, the grade point average earned in what were identified as solid subjects was the most valid one. In fact, among all the cognitive and non-cognitive variables this predictor was universally the most valid indicator of success in both the academic and the ward-performance phases of the training program, although certain part scores on the two *California* tests were the most valid predictors of standing in three of the courses.

3. In general, scales of the *MMPI* were not substantially predictive of success either in academic course achievement or in ward performance. Each of the four *MMPI* scales cited in Table 1 showed at least two statistically significant, though numerically low, validity coefficients (nearly all negative in sign) with level of achievement in selected courses. However, not one significant validity coefficient was found relative to the composite rating in ward adjustment.

4. As in previous studies, the relatively high intercorrelations among most of the pairings of courses pointed to the probable presence of some sort of "grade-getting syndrome," for which various hypotheses were set forth in a previous article by Michael, Haney, and Gershon (1963).

5. The median intercorrelation of .74 for the 12 items in the Ward Adjustment Rating Scale strongly suggested the possible presence of a halo factor and a possible tendency for raters to judge each trainee on all traits simultaneously rather than to evaluate a group of trainees on one trait at a time.

6. In light of the predictive validities reported for specific instruments employed in previous studies, the data of this study, which essentially furnished crossvalidation information for previously used predictors, point to the continued utility of the *California Reading Test*, the *California Mathematics Test*, and of the indices of high school achievement as the most promising predictors of success in

the academic program in nursing training. During the past several years only low predictive validities of the various scales of *MMPI* have been found relative to performance in ward activities. Indices of achievement in high school seem to hold at least modest promise for the prediction of this clinically oriented criterion variable.

Although the detailed results of three factor analyses are not presented, there were a few noteworthy findings. In support of the two hypotheses implied in statements four and five, two distinct factors with substantial loadings appeared—one in association with the 11 courses (variables 16–26) and the other in conjunction with the 12 items of the Ward Adjustment Rating Scale. In addition, three doublets, which were tentatively identified as verbal ability, mathematics ability, and spatial ability were respectively associated with substantial loadings on the vocabulary and comprehension portions of the *California Reading Test*, with high loadings on the parts of the *California Mathematics Test* concerned with reasoning and fundamentals, and with moderate loadings on the *EAST* No. 5—*Spatial Visualization* and the *SSRA* tests. Probably because of the existence of common items in more than one scale, the factors described by the intercorrelations of the scales on the *MMPI* were somewhat ambiguous and thus difficult to interpret.

Conclusion

Although the magnitudes of the validity coefficients were not high, the results of this investigation as well as of those of previous studies cited in the bibliography reveal that the *California Reading Test* and the *California Mathematics Test* as well as measures of high school achievement almost always have offered considerable promise in the prediction of the academic phases of the programs in nursing training at the Los Angeles County Hospital. Consistently, measures of spatial ability have failed to yield even modest predictive validities with respect either to standing in courses in the academic program or to the performance of ward activities. Results for the different scales from the *MMPI* instrument have varied considerably from study to study. At best, the predictive validities of these *MMPI* scales have been barely significant statistically, although the negative magnitudes of the coefficients have appeared with regular frequency from one validation study to the next. As was pointed out in a previous investigation by Michael, Haney, and

Gershon (1963), additional efforts need to be directed toward the improvement of the prediction of success in ward activities—especially through the use of autobiographical information which can be translated into operationally stated categories of observable behavior in the clinical work of nurses with patients. In short, it would appear that the existing battery of predictors with the exception of the spatial ability measures and the *MMPI* scales, can be substantially improved with refinements in the evaluation of nursing performance in the wards.

REFERENCES

- Haney, R., Michael, W. B., and Gershon, A. "Achievement, Aptitude, and Personality Measures as Predictors of Success in Nursing Training." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 389-392.
- Kaiser, H. F. "Computer Program for Varimax Rotation in Factor Analysis." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIX (1959), 413-420.
- Michael, W. B., Haney, R., and Gershon, A. "Intellective and Non-Intellective Predictors of Success in Nursing Training." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXXII (1963), 817-821.

CORRELATES OF ACHIEVEMENT ON THE ADMISSIONS TEST FOR GRADUATE STUDY IN BUSINESS

ARTHUR MITTMAN

University of Oregon

AND

JOHN W. LEWIS

Winona State College

Problem

ONE of the current problems in the education of graduate students in business administration is developing universal admissions criteria. As does the nature of the graduate programs vary, so do the criteria for admitting applicants. However, the *Admission Test for Graduate Study in Business* published by Educational Testing Service is rapidly being adopted as one objective measure of applicants' potential. Since this test has been adopted as a screening device, it seemed prudent to identify variables which account for performance on the test. Therefore, the purpose of this study was to investigate the relationship of relevant background variables to scores on the *Admission Test for Graduate Study in Business* for students enrolled in the Master of Business Administration program in the College of Business Administration at the State University of Iowa.

Variables

The background variables investigated were number of undergraduate semester hours in business, grade point average in undergraduate business courses, cumulative undergraduate grade point average, type of undergraduate college, and type of undergraduate

major. The criteria were the verbal and quantitative scores of the *Admission Test for Graduate Study in Business*.

Procedure

The types of undergraduate institutions from which the subjects matriculated were categorized as state universities in Iowa, private colleges in Iowa, state colleges or state universities outside of Iowa, or private colleges outside of Iowa. The types of undergraduate majors were categorized as business, social studies, biological science, physical science-mathematics, and engineering.

A stepwise multiple regression procedure was employed for investigating undergraduate semester hours in business, grade point average in undergraduate business courses, and cumulative undergraduate grade point average. This method allowed for testing the significance of each beta weight and for the deletion of those which did not reach significance ($\alpha = .05$). Simple analysis of variance was employed to test the null hypothesis of no differences in average test scores for type of undergraduate college and for type of undergraduate major.

Subjects

The subjects were 67 students enrolled in the Masters in Business Administration program at the State University of Iowa during the 1960-61 academic year.

Results

In Tables 1 and 2, respectively, are presented the data obtained from the multiple regression analysis of the relationship of number of undergraduate hours in business, grade point average in under-

TABLE 1

Multiple Correlation Coefficient (R), Standard Score Regression Weights (β), and Corresponding Test of Significance for the Regression Weights before and after Deletion of Nonsignificant Variables for Graduate Study in Business Verbal Test Score.

Statistic	Variable			
	No. of Hours	Bus. GPA	Cum. GPA	R
β	*.725			.66
β	*.697	.310	.175	.66
β	*.650	.388		.65

* $p < .05$.

TABLE 2

Multiple Correlation Coefficient (R), Standard Score Regression Weights (β), and Corresponding Test of Significance for the Regression Weight before and after Deletion of Nonsignificant Variables for Graduate Study in Business Quantitative Test Score (N = 67)

Statistic	Variable			
	Bus. GPA	Cum. GPA	No. of Hours	R
β	.144	.155	-.129	.26
β	.147	.181		.26
β	.232			.25

* $p < .05$.

graduate business courses, and cumulative grade point average as predictors with the verbal and quantitative scores of the *Admission Test for Graduate Study in Business* as criteria.

Number of undergraduate hours in business was the only background variable to yield a significant regression weight for predicting verbal test score. The correlation coefficient of .65 between number of undergraduate hours in business and verbal test score suggested a relatively high degree of relationship. None of the three background variables yielded significant regression weights for predicting the quantitative test score.

The analysis of variance on the verbal test scores did not yield significant F ratios for students classified according to type of undergraduate college ($F = 1.79$; $df = 3/63$) and major ($F = 1.54$; $df = 4/61$). The investigation of quantitative test score did yield significant results. The F ratios were 3.17 ($F_{.05} = 2.05$; $df = 4/61$) for type of undergraduate college and 4.47 ($F_{.05} = 2.76$; $df = 3/63$) for type of undergraduate major.

The multiple comparisons of quantitative score for students classified by type of undergraduate college and major are presented as

TABLE 3

Multiple Comparison for Graduate Study in Business Quantitative Test Scores for Students Classified by Type of Undergraduate College (N = 67)

College Category	Pri. Col. in Iowa	St. Col.—Univ. Outside of Iowa	Pri. Col. Outside Iowa
St. Univ. in Iowa	*6.43	*7.57	*5.32
Pri. Col. in Iowa		1.14	1.11
St. Col.—Univ. Outside Iowa			2.25

* $t_{.01} = 2.66$; $df = 60$.

TABLE 4

Multiple Comparisons on Graduate Study in Business Quantitative Test Scores for Students Classified by Type of Undergraduate Major (N = 67)

Major	Social Science	Bus.	Bio. Science	Phys. Sci. or Math
Engineer	*9.13	*6.23	2.13	1.25
Phys. Sci. or Math	*7.88	4.98	.88	
Bio. Science	7.00	4.10		
Business	2.90			

*t._{.05} = 2.00.

Tables 3 and 4. The number in each cell represents the difference between the group means.

The significant differences on quantitative score favored students from state universities in Iowa and students with engineering or physical science-mathematics undergraduate majors. A check of the data revealed that most of the students with strong quantitative backgrounds matriculated from state universities in Iowa. This fact may, in part, explain the differences obtained for type of undergraduate college.

Summary

The purpose was to investigate the relationship of relevant background variables to verbal and quantitative scores on the *Admission Test for Graduate Study in Business*. The subjects were 67 students enrolled in the MBA program at the State University of Iowa. The results justify the following general conclusions.

1. Number of undergraduate hours in business was the only background variable investigated which was significantly related to verbal test score.
2. Type of undergraduate college and major were both significantly related to quantitative score. Students from state universities in Iowa and students with engineering and physical science-mathematics undergraduate majors achieved significantly higher quantitative test scores.
3. Cumulative grade point average and grade point average in undergraduate business courses did not yield significant relationships with either verbal or quantitative test scores.

THE PEABODY PICTURE VOCABULARY TEST IN COMPARISON WITH OTHER INTELLIGENCE TESTS AND AN ACHIEVEMENT TEST IN A GROUP OF MENTALLY RETARDED BOYS¹

FRANCES M. THRONE, JOSEPH C. KASPAR

Children's Memorial Hospital

AND

JEROME L. SCHULMAN

Children's Memorial Hospital and Northwestern University Medical School
Chicago, Illinois

In the *Peabody Picture Vocabulary Test (PPVT) Manual*, Dunn (1959) reports "congruent validity" coefficients of .58 to .94 with other intelligence tests for the *PPVT* and "concurrent validity" coefficients of .39 to .87 with various achievement measures. These data were obtained in normal, retarded, and cerebral palsied groups. Kimbrell (1960) reports significant correlations between *PPVT* IQ's and *Wechsler Intelligence Scale for Children (WISC)* Verbal and Full Scale IQ's, and nonsignificant correlations with Performance IQ's for a sample of educable educational retardates. In the same study, the *PPVT* MA's did not correlate with grade placement on the *Gray-Votaw-Rogers General Achievement Tests*, while the derived *WISC* MA scores did. Tobias and Gorelick (1961) found significant correlations between the *PPVT* raw scores, the three IQ scores of the *Wechsler Adult Intelligence Scale (WAIS)*, the MA scores of the *Revised Stanford-Binet Intelligence Scale (S-B)*, specific form not given, and the reading subtest of the *Wide Range Achievement Test* for a group of adult retardates who were working in a community sheltered-workshop situation. However, the subjects scored

¹ This study was supported by a grant from the Lt. Joseph P. Kennedy, Jr., Foundation.

significantly higher on the *PPVT* than on either the *WAIS* or the *S-B*, which the authors attributed to greater opportunity for vocabulary growth provided by their setting. Budoff and Purseglove (1963) divided 46 institutionalized mentally retarded adolescents into two groups at an MA of 8-0. The lower MA group's *PPVT* MA scores correlated significantly higher with the *S-B* (L and L-M) MA scores than did the scores of the higher MA group. The only significant correlation for the higher MA group was between Form B of the *PPVT* and Form L of the *S-B*; the others were below significance. All the lower MA group's correlations were significant, however, as well as the correlations of the total sample. In addition, their *PPVT* scores were lower than their *S-B* scores on all forms. In discussing this latter finding which is in contrast to Tobias and Gorelick's (1961) finding, the authors suggested that institutionalized patients are minimally stimulated in the area of vocabulary, as opposed to Tobias and Gorelick's group of sheltered-workshop retardates.

The purpose of the present study was to investigate the *PPVT* as it is related to other intelligence tests and to an academic achievement test in another population of mentally retarded subjects, who were residential students in a school for exceptional children.

Method

The subjects were a group of 35 educable retarded boys, whose ages ranged from 11-0 to 14-11, and whose *S-B* (L and/or M) IQ's ranged from 50 to 77, and who resided at a school for retardates.² Through use of the Pearson product-moment correlational technique, their *PPVT* IQ and MA scores were correlated with other intelligence test scores from the *WISC*, the *S-B* (L and/or M), and the *Goodenough Draw A Person Test (DAP)*. In addition, the four test scores were compared with each other by t-test analysis.

In a factor-analytic study of the *WISC* using the standardization population, Cohen (1959) identified three factors for the 13 1/2 year old group: (1) *verbal comprehension*, composed of the Information, Comprehension, Similarities, and Vocabulary subtests; (2)

² The subjects were residents at the Lt. Joseph P. Kennedy, Jr., School for Exceptional Children located in Palos Park, Illinois. The assistance of the staff of that school is gratefully acknowledged.

perceptual organization, composed of the Block Design and Object Assembly subtests; and (3) *freedom from distractibility*, composed of the Arithmetic and Digit Span subtests. In the present study, each of these three *WISC* factors was represented by the average of the scaled scores of the component variables of the three respective factors. The resulting average scores were then correlated with the *PPVT* IQ and MA scores.

Finally, the *PPVT* IQ and MA scores were correlated with the average achievement scores obtained on the *Metropolitan Achievement Test (MAT)* by subjects, as well as with various individual subtests of the *MAT*: Word Discrimination, Word Knowledge, Average Reading, and Average Arithmetic.

Results

The correlations of the *PPVT* IQ's and *WISC* IQ's essentially corroborated Kimbrell's finding on subjects similar to the subjects of the sample employed in this sample: a significant relationship

TABLE 1
*PPVT Correlations with the WISC,
Stanford-Binet, and Goodenough Tests*

	PPVT IQ	PPVT MA
<i>WISC</i>		
Full Scale IQ	.45**	.70***
Full Scale MA	.42**	.45**
Verbal IQ	.53***	.72***
Performance IQ	.30	.25
Information	.34*	.33*
Comprehension	.49**	.46**
Arithmetic	.40*	.41*
Similarities	.34*	.35*
Vocabulary	.49**	.48**
Digit Span	.49**	.49**
Picture Completion	.13	.17
Picture Arrangement	.25	.19
Block Design	.28	.22
Object Assembly	.21	.14
Coding	.32*	.27
<i>Binet</i>		
IQ	.46**	.40*
MA	.34*	.39*
<i>Goodenough</i>		
IQ	.14	.05
MA	.01	.06

*Significant at .05 level.

**Significant at .01 level.

***Significant at .001 level.

obtained between the *PPVT* IQ's and the *WISC* Verbal and Full Scale IQ's ($r = .53, p < .001$; $r = .45, p < .01$, respectively). The correlation of the *PPVT* with the Performance Scale IQ was not significant. In addition, each Verbal Scale subtest scaled score correlated significantly ($r = .35$ to $.49, p < .05$ to $< .01$) with the *PPVT* IQ, whereas only one of the Performance Scale subtest scale scores (Coding) did the same. The *PPVT* MA correlations with the *WISC* IQ, MA, and subtest scale scores for the most part paralleled those of the *PPVT* IQ's; however, the *WISC* Full Scale and Verbal Scale IQ correlations with the *PPVT* MA's were somewhat higher than they were with the *PPVT* IQ's.

Both the *PPVT* MA's and IQ's correlated significantly with the *S-B* MA's and IQ's; however, nonsignificant correlations were obtained between the *PPVT* and *DAP* Scores (see Table 1).

The *t*-test comparisons indicated that of the four intelligence tests the *PPVT* yielded significantly higher scores than each of the other tests. The mean MA's (expressed in months) were: *PPVT*, 96.37; *S-B*, 89.80; *WISC*, 80.03; and *DAP*, 80.83. The *S-B* mean MA's were also significantly higher than the mean *WISC* MA's (see Table 2).

TABLE 2

t-Test Comparisons of the *PPVT*, *WISC*,
Stanford-Binet, and *Goodenough Tests'* MA's

	<i>PPVT</i> Mean = 96.37	<i>Binet</i> Mean = 89.80	<i>WISC</i> Mean = 80.03	<i>Goodenough</i> Mean = 80.83
<i>PPVT</i>	—	—		
<i>Binet</i>	2.37*	—		
<i>WISC</i>	5.20**	4.89**		
<i>Goodenough</i>	4.01**	1.54	.99	

*Significant at .05 level.

**Significant at .01 level.

Among the three *WISC* factors, the verbal comprehension and freedom from distractibility factors were correlated significantly ($r = .49$ to $.51, p < .01$) with the *PPVT* MA's and IQ's. The perceptual organization factor did not correlate with either of the *PPVT* sets of scores. These findings are not surprising in view of the subtests which make up the factors—*verbal comprehension* and *freedom from distractibility* being comprised of Verbal Scale sub-

tests, and perceptual organization of Performance Scale subtests (see Table 3).

TABLE 3
PPVT Correlations with Three WISC Factors

	PPVT IQ	PPVT MA
Verbal Comprehension	.51**	.51**
Perceptual Organization	.27	.19
Freedom from Distractibility	.49**	.50**

**Significant at .01 level.

Finally, again essentially corroborating Kimbrell's results regarding the validity of the *PPVT*, the average achievement scores did not correlate significantly with the *PPVT* IQ's or MA's (see Table 4).

TABLE 4
PPVT Correlations with Metropolitan Achievement Test

	PPVT IQ	PPVT MA
Over-all Achievement Average	.20	.29
Word Discrimination	.21	.79***
Word Knowledge	.17	.80
Average Reading	.54***	.27
Average Arithmetic	.28	.41*

*Significant at .05 level.

***Significant at .001 level.

According to these data, the *PPVT* MA's, although not related to over-all average achievement, were significantly related to the Word Discrimination and Arithmetic subtest scores ($r = .79$, $p < .001$; $r = .41$, $p < .05$, respectively). There was no relationship between MA and Word Knowledge and Average Reading; however, the *PPVT* IQ scores, on the other hand, did correlate significantly ($r = .54$, $p < .001$) with the Average Reading Scores—a finding which suggested that, for this population, the interaction of the MA and CA scores, not just the MA alone, was crucial to reading ability. That the MA correlated with Word Discrimination and not Word Knowledge is noteworthy in view of the differences between these two tests as compared to the *PPVT*. Actually, the Word Knowledge subtest is more similar to the *PPVT* in that for both tests the child is required to demonstrate his recognition and understanding of an orally presented word by correctly associating each word with a pic-

ture. That the correlation between these two tests is so low is surprising in view of their apparent similarities. The Word Discrimination subtest, on the other hand, which was highly correlated with *PPVT* MA ($r = .79, p < .001$) demands that the child select an orally presented word from a group of words of similar configuration in printed form.

Since none of the subjects in this study had *S-B* IQ's below 50, it was not possible to dichotomize the group of this study as Budoff and Purseglove did their group (IQ's above and below 50), in order to see whether differential results similar to theirs obtained. When the group was divided at its *S-B* MA mean (7-6), nonsignificant correlations were obtained between the *PPVT* and the *S-B*. However, the above-mean group's ($N = 16$) correlation was much higher (.40) than was the below-mean group's ($N = 19$) correlation (.08). These results were opposite from those found in Budoff and Purseglove's study, in which the correlation for the higher MA group with the *S-B* was much below that of the lower MA group. The groups are not entirely comparable, however, in that their sample not only included some subjects with lower IQ's (three with IQ's less than 20) but also included older subjects (16 to 18 years old).

Conclusion

Just as the results of previous research have shown, those of the present study indicate that the *PPVT* demonstrates a high degree of "congruent" validity, when other intelligence tests are used as criterion measures.

With reference to the question of the *PPVT* scores being depressed because of the subject's institutionalization, the subjects in this study whose *PPVT* scores were considerably higher than their corresponding scores on the other intelligence tests (similar to Tobias and Gorelick's results) were all residential students. This circumstance tends to suggest that institutionalization *per se* does not depress verbal scores, but the kind of institution (i.e., verbally stimulating or not) is the crucial factor. The subjects in the current study were residents of a private school, whereas Tobias and Gorelick's were day-workers in a community sheltered-workshop; Budoff and Purseglove's subjects, on the other hand, were inmates of a state institution.

It should be noted that the *PPVT* has not been shown to serve as an adequate predictor of academic achievement either in this study or in Kimbrell's. This finding is true in spite of the fact that the *PPVT* correlates significantly with other intelligence tests which in turn correlate significantly with achievement tests.

Summary

The *PPVT* results of 35 mentally retarded boys were compared with their *WISC*, *S-B*, and *DAP* scores as well as with their overall average achievement scores and certain subtest scores of the *MAT*. Significant correlations were obtained between the *PPVT* and all the intelligence tests except the *DAP*; the *PPVT* yielded significantly higher scores in all instances. In addition, two of the three *WISC* factors (Cohen, 1959) correlated significantly with the *PPVT*. When compared with the average achievement scores of the *MAT*, no predictability obtained; but various subtests (Word Discrimination and Arithmetic) were related to *PPVT* MA. These findings point to the necessity for guarding against assuming that intelligence tests correlated with each other necessarily predict each other's achievement-test correlates too.

REFERENCES

- Budoff, M. and Purseglove, Eleanor M. "Peabody Picture Vocabulary Test Performance of Institutionalized Mentally Retarded Adolescents." *American Journal of Mental Deficiency*, LXVII (1963), 756-764.
- Cohen, J. "Factorial Structure of *WISC* at Ages 7-8, 10-8, and 13-6." *Journal of Consulting Psychology*, XXIII (1959), 285-299.
- Dunn, L. M. *Manual, Peabody Picture Vocabulary Test*. Nashville: American Guidance Service, 1959.
- Kimbrell, D. L. "Comparison of Peabody, *WISC*, and Academic Achievement Scores among Educable Mental Defectives." *Psychological Reports*, VII (1960), 502.
- Tobias, J. and Gorelick, J. "The Validity of the Peabody Picture Vocabulary Test as a Measure of Intelligence in Retarded Adults." *Vineland Training School Bulletin*, LVIII (1961), 92-98.

PMA FACTORS, SEX, AND TEACHER NOMINATION IN SCREENING KINDERGARTEN GIFTED

PHILLIP WEISE

Pasadena City Schools

C. E. MEYERS

University of Southern California

■ ■ ■

JOHN K. TUEL

University of California at Los Angeles

THIS study was concerned with the efficiency of nomination of kindergarten children for expensive *Binet* testing when they are nominated for placement in special programs for the gifted. It was important to avoid excessive fruitless testing. In the initiation of the present study, 70 percent of teacher-nominated children did reach the *Binet* IQ criterion of 130. Of the "misses," some were below 100. On the other hand, it was important to minimize the other error, that of missing non-nominated children who otherwise qualified for the program.

To reduce both risks was the purpose of this study. Another objective was to show a methodological model for similar work. The study was not concerned either with the ultimate validity of the *Binet* criterion or with the pros and cons of identifying and programming at the kindergarten level. Accepting these practices, the issue was to make the selection for individual testing more efficient by adding to the teacher nomination the information gained by (a) group factor tests and (b) the boy-girl differences in them.

Subjects and Procedures

Subjects were drawn from a pool of 470 kindergarten children in 19 classes under 10 teachers. The classes were so selected as not

only to be representative of all such classes in the large school district in question, but also to utilize only reasonably experienced kindergarten teachers. Initial selection of nominees for individual testing was made by teachers, who were free to name as few or as many as they believed might earn the privilege of special programming. The teachers nominated 57.

The 1953 form of the *SRA Primary Mental Abilities* for ages 5 to 7 (*PMA*, 5-7) was administered to the total pool of 470 children in small groups of 3 to 5. Only the Verbal (*V*), Perceptual (*P*), Quantitative (*Q*), and Spatial (*S*) subtests were employed, the Motor measure being omitted. The *Binet* was then administered to all children nominated by teachers, as well as to those children who received high scores on the *PMA*. As experience cumulated, cut-offs were successively approximated to increase the efficiency of selection. It was found that differences between boys and girls could be utilized, and combinations of teacher nomination and test scores could be applied. Eventually, combinations of teacher nomination and test scores were developed to eliminate the time and expense of screening children who would almost certainly not achieve the *Binet* criterion.

Findings

In all, 93 children were given the *Revised Stanford-Binet Intelligence Scale*, Form L-M; however, complete data for statistical

TABLE 1
Frequencies and Percentages of Teacher Nominated and Non-nominated Children Attaining and Not Attaining Binet IQ 120

IQ > 120	Boys		Girls		Totals	
	Number	Percent	Number	Percent	Number	Percent
Nominated	18	43.0	17	28.5	35	40.5
Not Nominated	7	16.5	9	20.5	16	18.5
		(28.0) ^a		(34.6) ^a		
Totals	25	59.5	26	59.0	51	59.0
IQ < 120	Boys		Girls		Totals	
	Number	Percent	Number	Percent	Number	Percent
Nominated	10	24.0	8	11.5	18	17.5
Not Nominated	7	16.5	12	29.5	20	23.5
Totals	17	40.5	18	41.0	35	41.0
Grand Totals	42	100.0	44	100.0	86	100.0

^aPercentages are shown for each column independently.
^bPercentages of 26 and 28.

treatment were available for only 86. Table 1 indicates the *Binet* results, shown not only for the nominees, but also for the non-nominated remainder who were tested. As a whole, these ten teachers were accurate, but the table clearly shows this accuracy was diminished by the fact that 28 percent of the gifted boys and 35 percent of the gifted girls had not been nominated by these same teachers. Not evident in the tables are some IQ's over 160, especially in boys, in those children not nominated by teachers.

Table 2 presents multiple regression predictions of *Binet* IQ in two ways: (a) as a dichotomy of IQ under 130 vs 130 or higher, and (b) as a variate. Independent variables are teacher nomination

TABLE 2

Means, b-weights, and Other Data, Predicting Binet IQ As a Dichotomy or a Variate

	Binet IQ as Dichotomy (129-130)		Binet IQ as Variate	
	Mean	b-wt	Mean	b-wt
Boys N = 42				
Teacher Nomination	.667	.279	.667	12.071
PMA-V	115.976	.011	115.976	.029
PMA-P	119.738	-.002	119.738	-.191
PMA-Q	119.167	.006	119.167	.608
PMA-S	107.024	.006	107.024	.273
Binet IQ	.595		134.301	
Intercept		-1.997		-12.168
S.E._est.		1.140		17.176
Girls N = 44				
Teacher Nomination	.500	.606	.500	7.679
PMA-V	115.091	.021	115.091	.301
PMA-P	130.318	.014	130.318	3.16
PMA-Q	120.364	-.003	120.364	.790
PMA-S	108.795	.001	108.795	.125
Binet IQ	.568		120.418	
Intercept		-3.641		-10.623
S.E._est.		.437		11.601
Total Group N = 86				
Teacher Nomination	.581	.309	.581	8.991
PMA-V	115.523	.016	115.523	.576
PMA-P	125.151	.004	125.151	.051
PMA-Q	119.779	.001	119.779	.213
PMA-S	107.930	.003	107.930	.196
Binet IQ	.581		122.558	
Intercept		-2.296		-9.114
S.E._est.		.459		12.156

as gifted (Yes or No) and the four *PMA* subtest IQ's, *V*, *P*, *Q*, and *S*. It is readily seen that predictions differ between the sexes. Teacher nomination and the *V* subtests are useful for both sexes, while accuracy of prediction is increased by use of *Q* for boys and of *P* for girls. Consistently, one finds negative or low "b" weights for *P* in boys and for *Q* in girls. It appears, then, that time and effort can be saved by limiting subtest screening to *V* and *Q* for boys and to *V* and *P* for girls. To help implement this recommendation, Figures 1 to 4 were constructed from the multiple regression data. From these graphs a prediction of *Binet* test results may be made by "plugging in" the teacher nomination (Yes or No), the *V* subtest score, and either the *Q* subtest score for boys or the *P* subtest score for girls. In Figure 1, for example, the probability of a boy's reach-

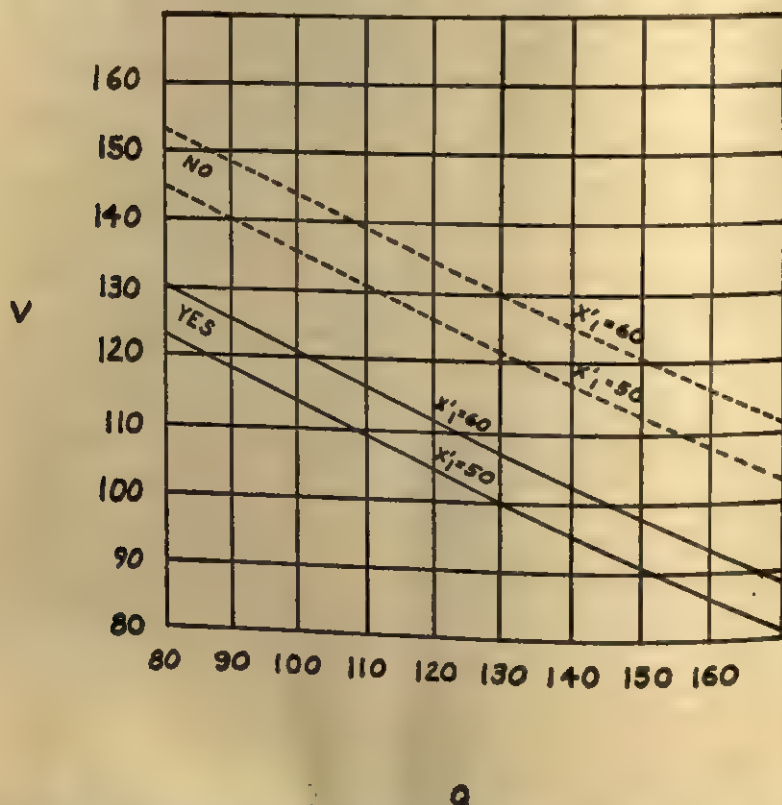


Figure 1. Probability of teacher nominated and non-nominated boys attaining *Binet* IQ 130 or greater with any combination of *V* and *Q* subtest scores on the *PMA*. X'_1 = probability of attaining criterion. "Yes" means teacher nomination. "No" means teacher non-nomination.

ing or surpassing the *Binet* criterion (130) with a V of 110 and a Q of 120 is .60 if he was nominated as gifted (Yes) by his teacher. For a boy to have a .60 chance of reaching criterion *without* teacher recommendation (No), he would need to attain, for example, a V of 132 and a Q of 120 or any other combination of intersecting V and Q scores along the $X = .60$ (No) line of Figure 1. Figure 2 presents

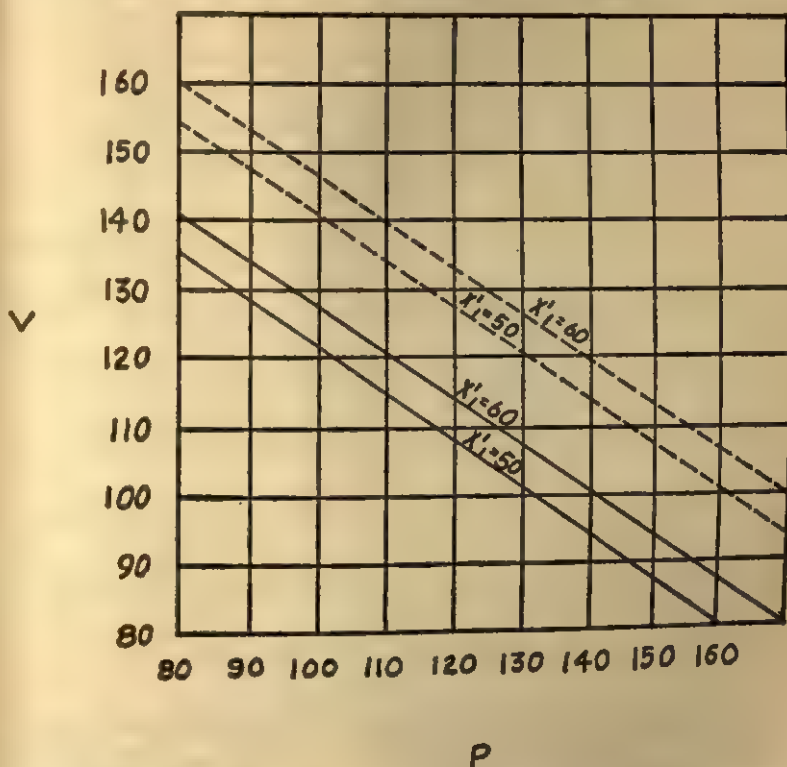


Figure 2. Probability of teacher nominated and non-nominated girls attaining *Binet* IQ 130 or greater with any combination of V and P subtest scores on the *PMA*. X'_1 = probability of attaining criterion. "Yes" means teacher nomination. "No" means teacher non-nomination.

similar information for girls when the V and P variables are used. Figures 3 and 4 show the predicted *Binet* IQ for teacher nominated and non-nominated boys and girls. A nominated boy with a V of 120 and a Q of 116 would thus be predicted to achieve a *Binet* IQ of 140, whereas a nominated girl with a V of 120 would have to have a P of 137 to attain 140 on the *Binet*.

Thus, within a reasonable range of possible *PMA* scores, these two tables can be used to predict probable *Binet* scores with and without teacher nomination on the basis of just two *PMA* subtest scores. Limitations on the numerical value of the *PMA* quotients impose practical limitations on the extension of these tables for a probability of .80 or greater or in predicting *Binet* quotients beyond 150. For example, the maximum possible *PMA* quotient for CA 5-3 is 171, decreasing to 140 high *Binet* IQ. Apparently girls have a variance in the *P* subtest useful for anticipating *Binet*—a variance not found in boys, who in contrast appear to vary usefully in *Q*. Whatever theory may eventually explain the findings, it is apparent that efficiency can be served by exploitation of the differ-

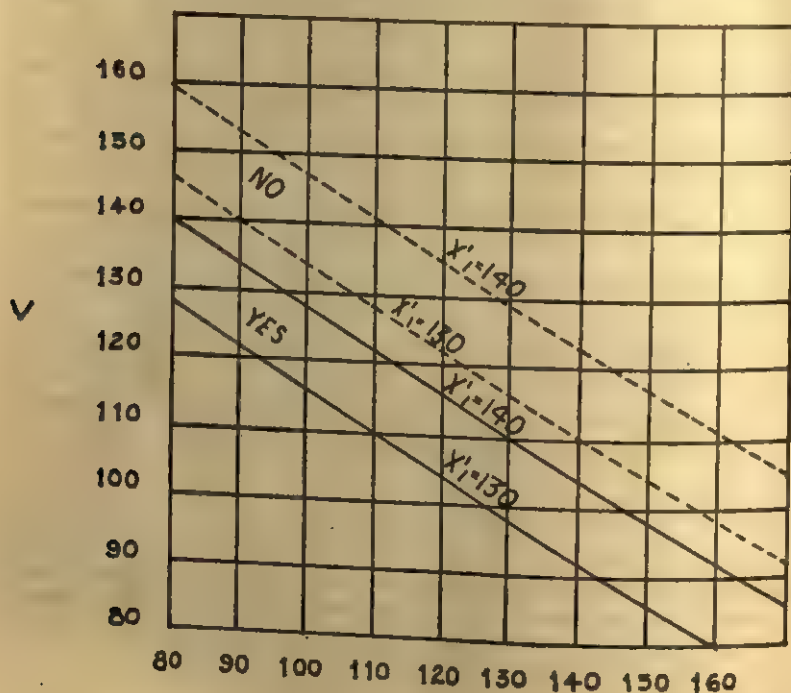


Figure 3. Predicted *Binet* IQ for teacher nominated and non-nominated boys for any combination of *V* and *Q* subtest scores on the *PMA*. X'_1 = predicted *Binet* IQ. "Yes" means teacher nomination. "No" means teacher non-nomination.

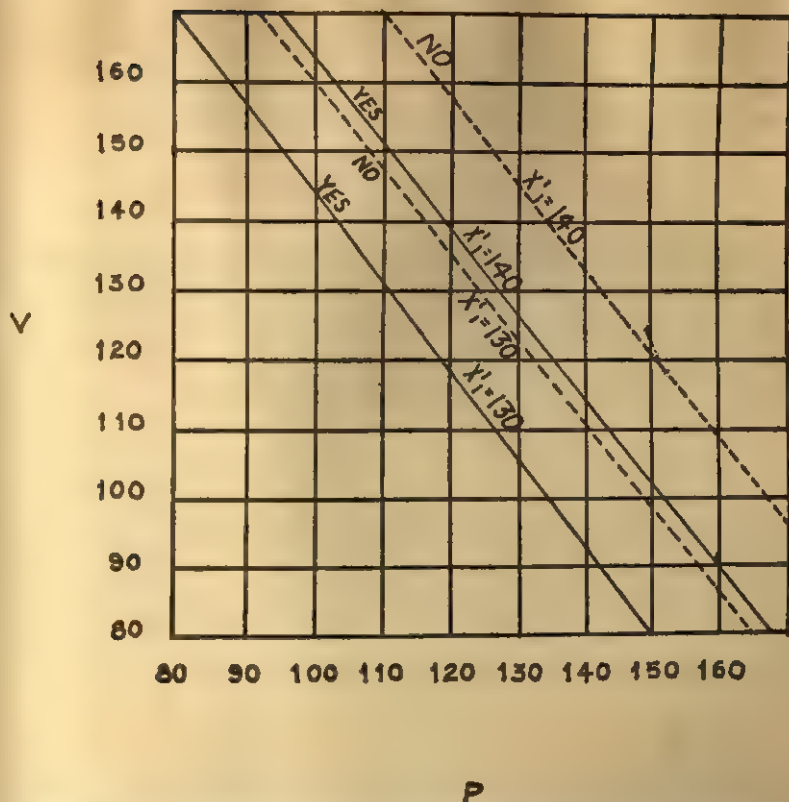


Figure 4. Predicted *Binet* IQ for teacher nominated and non-nominated girls for any combination of *V* and *P* subtest scores on the *PMA*. X_1 = predicted *Binet* IQ. "Yes" means teacher nomination. "No" means teacher non-nomination.

ence. This sex differential prediction obtains either with or without teacher nomination. It is noted in passing that teacher nomination alone is itself of considerable predictive value. It is also noteworthy that the subtest weights given in the *PMA Manual* (Thurstone and Thurstone, 1953) used to construct the profile scales may be misleading, at least for the range of intelligence involved in the present context, since no sex distinction was made.

REFERENCE

Thurstone, L. L. and Thurstone, T. G. *Technical Supplement for P.M.A. for Ages 5 to 7*. Chicago: Science Research Associates, Inc., 1953.

THE APPLICATION OF A CONFIGURATION METHOD TO THE PREDICTION OF SUCCESS IN FIRST GRADE

JEAN L. BALINKY

Bridgewater-Raritan School District

THE recent tendency toward a proliferation of ungraded or semi-graded primary school programs has been symptomatic of the growing realization that the traditional chronological age criterion is inadequate as the sole determinant of readiness for a formal academic program.

The principal source of difficulty in the establishment of any of these programs lies in the assignment of pupils to various groups (Mattick, 1963). The standardized reading readiness tests, intelligence tests, and teacher evaluations have all been used as predictive devices with varying success.

High scores on reading readiness tests, for example, are a good indicator of potential success; low scores are not an equally good predictor of failure (Lee, Clark, and Lee, 1934). Intelligence test scores have been found to have a positive correlation with early school achievement, but are still inaccurate in a high proportion of cases (Morgan, 1960). Teacher judgments of potential success or failure vary radically from one rater to another. On the whole, teachers have been found to err on the side of over-prediction (Mattick, 1963).

Individualized methods of evaluation, in which a variety of measures including verbal ability, visual-motor coordination, abstract abilities, and maturational level are utilized, have been found to be effective (DeHirsch, 1957; Ilg, 1963). Their use is limited by lack of personnel to carry out such evaluations within most public school systems. Thus the necessity for developing a predictive method in

which group techniques are utilized and in which a minimum number of specially trained personnel is required becomes apparent.

The concept of instructional grouping implies the prediction of certain discrete categories of outcome rather than any kind of ordinal result. Therefore, limiting prediction to two or three possible categories of outcome affords a means of obtaining very accurate predictive results.

In most methods of prediction, linearity of data, or at least the assumption of linearity, is necessary. Valuable information may be lost because it cannot be incorporated within the statistical design. Information may be distorted by the imposition of statistical assumptions of equality of units or dimensionality. Further inaccuracy may result if it is assumed that all predictive measures apply to all cases within the sample. The present study applies Stuckert's (1958) method of prediction by configuration to the problem of determining potential success in first grade. This method was selected because it eliminates most of these difficulties and is readily applicable in a situation where computational equipment is usually unavailable.

The Configurational Method

This technique permits the prediction of discrete categories of outcome based on performance on one or more predictor variables, continuous or discrete, on the basis of the principle of maximum probability. It is based on the concept that in any large sample there will be relatively homogeneous subsamples that will tend to have a high probability of having a similar behavioral outcome. If the original large sample can be divided into subsamples based on homogeneity of several factors, then the behavioral outcome should be predictable. Further, other cases not included in the original subsample, yet having a similar rating on each of the factors, should tend to behave in a similar manner. If similar subsamples can thus be derived on the basis of a population with known outcome, it should be possible to predict the behavior of future cases.

The level of predictive accuracy is chosen arbitrarily. If it is too low, the prediction is of little value; if too high, it is not possible to find enough homogeneous subsamples. Thus too large a residual of unpredictable cases is left. In practice, an arbitrary cutoff point termed a *critical value* is chosen. This *critical value* corresponds to a

certain *proportion of cases* which must fall into a given outcome category before that category is considered predictable. Ideally the *critical value* should be the same for every subsample. However, in some situations it may be necessary to alter the value in order that some kind of prediction may be effected.

Selection of the initial predictor variable is made on the basis of the variable having the greatest number of cases in excess of the *critical value*. Cases exceeding the *critical value* on the initial variable require no further treatment. The remaining cases are divided into subsamples on the basis of additional predictor variables.

Procedure

Subjects were children enrolled in the kindergartens of the Bridgewater-Raritan Schools during the 1961-62 and 1962-63 academic years. All pupils had to be five years of age by December first of the kindergarten year. Pupils attending kindergarten during the 1962-63 school year were designated the *criterion sample*, whereas those in the 1961-62 class were considered to be the *validation sample*.

Predictive data were collected from March through May of the kindergarten year. The final choice for a criterion measure was the instructional reading level in January of the first-grade year. Inter-teacher variability was reduced by the reading supervisor who kept careful checks on the level of instruction in each classroom.

Two types of prediction were attempted. The first was a simple success-failure dichotomy, with success defined as reading in a primer by January; failure, as anything below that level. A three-category prediction, "success-doubtful-failure" was also tried. Here the "success" criterion remained the same; "doubtful" was defined as reading in a third preprimer or transition text; and failure was anything below that level.

The ultimate selection of a two- or three-category prediction is dependent upon the objective of the predictive work. In the present study, consideration of the nature of the school program was the essential factor. Within the confines of the traditional graded program, a simple success-failure dichotomy was sufficient. However, had some type of ungraded primary program been available, there would have been an interest in differentiation of those pupils who could profit from a modified academic program as apart from those

who were destined for total failure. In the latter situation, a tripartite result would seem to be highly desirable.

The best results in configurational prediction tend to be obtained when the widest variety of predictive data is available. There is no assumption that all predictors are equally significant for all subsamples. In some subsamples, only one or two predictors may be necessary, whereas for other groups a large number of predictors may be required for accurate results.

The selection of predictive devices was based in large measure on clinical evidence of causes of learning disability in children. Thus, factors such as intelligence, motor coordination, and vision (DeHirsch, 1957) were considered to be relevant predictors. These were combined with teacher judgment and teacher identity (Kermoian, 1962). The particular population in this study was a fairly homogeneous group in terms of socioeconomic level. (The pupils came from middle class homes in which most parents had had at least 12 years of formal education.) For this reason, no measures of either economic or educational level of the parents were included. However, should prediction be attempted in a more heterogeneous community, some measures of these factors might be incorporated. The predictors employed are cited in Table 1.

TABLE 1
Summary of Predictors Used in the Configurations

Two Category Prediction	Three Category Prediction
Goodenough <i>Draw-a-Man</i> Mental Age (Goodenough, 1926)	Goodenough <i>Draw-a-Man</i> Mental Age (Goodenough, 1926)
Starr <i>Rutgers Drawing Test</i> (Starr, 1952)	Starr <i>Rutgers Drawing Test</i> (Starr, 1952)
<i>Metropolitan Readiness Tests</i> (Hildreth and Griffiths, 1949)	<i>Metropolitan Readiness Tests</i> (Hildreth and Griffiths, 1949)
Total Score	Total Score
Verbal Score	Verbal Score
Arithmetic Score	Arithmetic Score
Copying Score	Copying Score
Kindergarten Rating—3 point scale	Kindergarten Rating—3 point scale
Identity of Kindergarten Teacher	Identity of Kindergarten Teacher
Identity of First Grade Teacher	Identity of First Grade Teacher
Month of Birth	Month of Birth
	Sex

As each set of data was collected, it was punched into IBM pre-punched cards. This method was selected because it eliminated the

need for compiling long lists of combined data. The completed cards were mechanically reproduced for use in the sorter.

On the basis of a pilot study in which a portion of the 1961-62 sample was utilized, the mental age score from the Goodenough's (1926) *Draw-A-Man* Test was chosen as the initial predictor because it showed marked variability in level that correlated positively with reading achievement. Cut-off points for the configurations were determined on the basis of preliminary scatter diagrams. A critical value of .90 was chosen.

Configurations were constructed for the entire criterion sample. No attempt was made to predict outcome for those subsamples which did not exceed a single critical value, although any difference above the .50 level could have been utilized. The configurations were then applied to the prediction of cases in two validation samples. The first validation sample consisted of cases from the schools where the criterion sample was obtained. The second validation sample was composed of pupils from another school within the district where the population differed in ethnic background, socioeconomic level, and parental education background.

TABLE 2
Summary of Results

Information About Predictions	Criterion Sample (N = 303)		Validation Sample I (N = 289)		Validation Sample II (N = 59)	
	2-way pre- dictions	3-way pre- dictions	2-way pre- dictions	3-way pre- dictions	2-way pre- dictions	3-way pre- dictions
Percentage predicted (exceeding critical value of .90)	88	84	88	76	76	76
Percentage predicted correctly	78	76	74	62	51	53
Coefficient of efficiency*	.22	.15	.35	.03	.24	.26

*The proportion of reduction in error that is attained through the use of the instrument in relation to the use of no instrument.

Results

As the figures in Table 2 indicate, the accuracy of prediction is higher in the sample more closely resembling the criterion sample. A significant drop in the accuracy of prediction occurs when the

configurations are applied to a sample drawn from a different population. In the first validation sample, there is a significant difference in predictive accuracy between the two- and three-category predictions. Although the results for the second validation sample do not follow this pattern, the difference in proportions is not significant.

Discussion

There has never been a great deal of difficulty in identifying those pupils who will do exceptionally good academic work. The principal difficulties have arisen in attempting to forecast success for those who score within the average or below average range on single predictive devices. The value of the configurational method may lie in the fact that it can take into account some compensatory mechanisms which may enable certain pupils to succeed despite low scores on the more global predictive measures.

The possibility of incorporating the teacher as a predictor variable is one of the more important features of the method. It permits a relatively objective method of assessing a particular teacher's success with one or another type of pupil. There are also implications for the evaluation of teachers whose pupils consistently perform below predicted levels.

The emphasis throughout this study has been upon the method of prediction rather than upon the particular predictor values, which are applicable only to the school situation where they were developed. The method of predicting by configuration is, however, readily adaptable to any school situation as well as to a variety of other prediction problems. Adjustment of criterion values and substitution of a variety of predictive and criterion measures is easily accomplished. The selection of predictive devices is determined by a) the nature of the possibly relevant variables, b) the availability of test materials, and c) the presence of personnel for the administration of screening devices.

The key steps in the preparation of configurations may be summarized as follows.

1. Select the *critical value*.
2. Prepare scatter diagrams showing the relationship of each predictor to the criterion variable.

3. Select that predictor variable which differentiates to the highest degree the various portions of the sample.
4. Divide the sample into subsamples on the basis of the level of performance on the first predictor.
5. Calculate the proportions of each outcome category for each subsample.
6. Eliminate all cases which have exceeded the critical value.
7. Repeat steps 4 through 6 with the second and all succeeding predictors. Continue this process until the predictors have been exhausted, until all cases have been predicted, or until the number of cases within a given subsample becomes too small for predictive purposes.

Summary

The Stuckert method of prediction by configuration was applied to the prediction of success in first grade. A variety of predictive devices, including measures of intelligence, verbal ability, visual-motor coordination, age, sex, teacher identity, and teacher ratings of potential success were employed. The criterion was instructional reading level. Prediction of a success-failure dichotomy was accurate for 74 percent of the 289 cases; three-category prediction of "success," "doubtful," and "failure" was accurate for 62 percent of cases from the same population. The advantages of this predictive system over single predictive devices was discussed, as was its advantage over other multiple factor predictive methods which require the imposition of statistical assumptions and the use of elaborate computational equipment.

REFERENCES

- DeHirsch, K. "Tests Designed to Discover Potential Reading Difficulties at the 6-Year Old Level" *American Journal of Orthopsychiatry*, XXVII (1957), 566-576.
- Goodenough, F. L. *The Measurement of Intelligence by Drawings* Yonkers-on-Hudson, New York: World Book Company, 1926.
- Hildreth, G. H. and Griffiths, N. L. *Metropolitan Readiness Tests* New York: Harcourt, Brace & Company, 1949.
- Ilg, F. L. Personal communication, September, 1963.
- Kernolan, S. B. "Teacher Appraisal of First Grade Readiness" *Elementary English*, XXXIX (1962), 196-201.
- Lee, J. M., Clark, W. W., and Lee, D. M. "Measuring Reading Readiness." *Elementary School Journal*, XLIV (1934), 650-666.

- Mattick, W. E. "Predicting Success in the First Grade." *Elementary School Journal*, LXIII (1963), 273-276.
- Morgan, E. F. "Efficiency of Two Tests in Differentiating Potentially Low from Average and High First Grade Achievers." *Journal of Educational Research*, LIII (1960), 300-305.
- Starr, A. S. "The Rutgers Drawing Test." *The Training School Bulletin*, XLIX (1952), 45-64.
- Stuckert, R. P. "A Configurational Approach to Prediction." *Sociometry*, XXI (1958), 225-237.

THE PREDICTIVE VALUE OF A BEGINNING FIRST-GRADE INTELLIGENCE EXAMINATION

CARRIE M. SCOTT

Bend Public Schools

Problem

EDUCATORS are interested in the use of a measurement which will predict the academic achievement of a pupil upon entrance to the first grade. To accomplish this end, objective tests have been devised to measure the pupil's status and to predict academic success.

Since intelligence tests are administered at the end of kindergarten or at the time of first-grade entrance, and since the results are considered to be of significant value in indicating later school success, the writer investigated the relationship existing between the total scores obtained on the *Detroit Beginning First-Grade Intelligence Examination* and the grade-scores on the *Stanford Achievement Test, Primary Series* subtests.

Statistical Procedure

For the first part of this investigation, the intelligence test scores were correlated with grade-scores obtained on each subtest of the achievement test for 905 pupils in grade 2.8 of the elementary schools of Bend, Oregon. The second part deals with the study of 15 of the 905 pupils who rated average or superior in scores on the beginning intelligence examination, but did not achieve as expected. The total scores obtained on the *Detroit Beginning First-Grade Intelligence* test were compared with the results of later intelligence examinations, and with achievement test scores concerning the amount of retardation. In addition, consideration was given to special treatment of the 15 students previously mentioned, by the

Department of Special Education. Comparisons were made between grade-scores obtained on grade 2.8 level and 4.8 level on the *Stanford Achievement Test*.

Statistical Analysis

When the total scores obtained on the *Detroit Beginning First-Grade Intelligence Examination* were correlated with grade-scores obtained on the subtests of the *Stanford Achievement Tests, Primary Series* administered at grade 2.8 level, the range of total scores on the intelligence examination was from 25 to 105. The mean score was 70.13, the mean mental age was 6 years 10 months, and the mean IQ was estimated to fall in an interval from 96 to 104, a rating of C.

In Table 1 the range in grade-scores on the subtests of the achieve-

TABLE 1

Ranges, Means, and Standard Deviations of the Scores on the Sub-tests of the Stanford Achievement Test Administered in Grade 2.8

Subtest	Range in Grade Scores	Mean	Standard Deviation
Paragraph Meaning	1.3-7.5	3.6	.384
Word Meaning	1.3-6.6	3.3	.294
Spelling	1.0-5.3	3.5	.268
Arithmetic Reasoning	1.0-5.1	3.2	.253
Arithmetic			
Computation	1.0-4.8	2.9	.113
National Norm		2.8	

ment test is shown together with their means and probable errors. The widest range of grade-scores was in Paragraph Meaning, followed in turn by Word Meaning, Spelling, Arithmetic Reasoning, and Arithmetic Computation. All means were above the norm, grade 2.8.

Total scores on the intelligence examination were correlated with grade-scores on each subtest of the achievement test. The correlations along with their probable errors are shown in Table 2.

All correlation coefficients were significant beyond the .01 level. The highest coefficient of correlation was between intelligence test scores and scores on the test of Arithmetic Reasoning followed in turn by coefficients of correlation between intelligence test scores and the subtests of Paragraph Meaning, Word Meaning, Spelling,

TABLE 2

Correlations between Total Scores on the Detroit Beginning First-Grade Intelligence Examination and Scores on the Subtests of the Stanford Achievement Test, Primary Series, Administered in Grade 2.8

Test	Correlation	P.E.
Paragraph Meaning	.48	.016
Word Meaning	.43	.018
Spelling	.36	.019
Arithmetic Reasoning	.54	.016
Arithmetic Computation	.35	.018

and Arithmetic Computation. Therefore, positive relationships existed between the total scores obtained on the intelligence examination administered on grade 1.0 level and the subtests of the achievement test administered at grade 2.8 level.

Fifteen of the 905 pupils were selected who had rated average or superior on the *Detroit Beginning First-Grade Intelligence Examination*, and who had obtained mean reading grade-scores below grade 2.8 on the *Stanford Achievement* subtests in Paragraph Meaning and Word Meaning. The total scores on the beginning intelligence examination were compared with intelligence quotients on intelligence tests administered later, namely: the *Stanford-Binet Intelligence Scale*, the *Wechsler Intelligence Scale for Children*, and the *California Test of Mental Maturity*. The comparisons appear in Table 3. With the exception of Pupil E, the ratings on the subsequent intelligence examinations were average or superior and did not coincide with the results obtained on the achievement test.

Fourteen of the 15 pupils received aid from the Department of Special Education in the form of speech therapy and/or remedial reading. This aid factor may account for the mean retardation in reading being less than the mean retardation on the entire test battery in grade 4.8. Eleven of the pupils who received aid in reading were reported as improving, while four were reported to have made little gain in reading. Three pupils received both speech therapy and remedial reading. Since teachers referred these pupils for remedial aid, they were aware that they were not achieving as should be expected.

Mean grade-scores earned on the *Stanford Achievement Tests, Primary Series* administered in grade 2.8 and earned on the *Elementary Series*, administered in grade 4.8 were compared according

TABLE 3

*Comparison of Intelligence Quotients Attained by Fifteen Pupils
Whose Achievement Test Results Did Not Support the Total Scores
Received on the Detroit Beginning First-Grade Intelligence Examination*

Pupil	Detroit Beginning First-Grade Intelligence Examination	Stanford- Binet Intelligence Scale	Wechsler Intelligence Scale for Children		California Test of Mental Maturity	
			V.	P.	L.	N.L.
A	111-117		100	111	105	125
B	111-117		89	99		
C	118-Over	100				
D	96-104	99			86	117
E	96-104	102	87	80	113	85
F	96-104	100			101	101
G	118-Over	111			125	103
H	96-104	110			130	94
I	111-117	96				
J	96-104				106	85
K	96-104	105			113	124
L	105-110	113			98	94
M	111-117				108	100
N	118-Over				125	99
O	105-110	93	90	106	90	94

V—Verbal Scale

P—Performance Scale

L—Language

N.L.—Non-language

to mean grade-scores on the battery and mean reading grade-scores. On grade 4.8 level, the mean retardation had increased on the battery, but had decreased when mean reading grade-scores were considered. These data are presented in Table 4. *Example:* on grade 2.8 level, pupil A rated a mean grade-score of 2.7 on the achievement test, a retardation of .1 grade, and a mean reading grade score of 1.4 representing a retardation of 1.4 grades. On grade 4.8 level he rated a mean-grade-score of 3.4, a retardation of 1.4 grades on the battery, and a mean reading grade-score of 3.3, representing a retardation of 1.5 grades.

Summary

1. A wide range existed in grade-scores on the *Detroit Beginning First-Grade Intelligence Examination* administered upon first grade entrance.
2. The range in grade-scores on the *Stanford Achievement Test* administered in grade 2.8 was also wide.
3. That the correlations between intelligence examination scores

TABLE 4

Grade-Scores of Fifteen Selected Pupils in Grades 2.8 and 4.8 According to Results on the Stanford Achievement Test in Terms of Means on the Entire Test Battery and Mean Reading Grade

Pupil	Grade 2.8		Grade 4.8	
	Test Mean	Reading Mean	Test Mean	Reading Mean
A	2.7	1.4	3.4	3.3
B	2.3	2.0	*	*
C	2.5	2.4	3.7	3.6
D	2.3	2.5	4.0	4.2
E	2.0	1.9	4.8	4.8
F	2.6	2.4	4.4	4.2
G	2.6	2.3	4.1	3.8
H	2.2	1.8	3.1	3.2
I	2.3	2.5	5.9	6.2
J	2.7	2.2	4.5	4.6
K	2.6	2.3	4.4	4.7
L	1.9	1.9	3.3	4.7
M	2.7	2.9	3.8	4.2
N	2.1	2.0	4.5	5.4
O	1.9	1.9	4.4	4.3
Mean	2.36	2.16	4.16	4.30

*Data unavailable.

and scores on the subtests of the achievement test were positive, indicated that a relationship existed, as might be expected with cognitive measures.

4. Fourteen of the 15 pupils selected who were average or above on the beginning intelligence examination placed similarly on intelligence examinations administered later.
5. Fourteen of the 15 pupils who were academically retarded at grade 2.8 level continued to be retarded on grade 4.8 level.
6. There was less reading retardation by the end of grade four, probably because of the additional aid administered by the Department of Special Education.

Conclusion

The results of this investigation indicated that school success cannot be predicted from mental tests alone but that many other contributing factors operate. Since teaching is geared to the so-called average child, learning does not proceed according to the potentiality of every child. However, a beginning intelligence examination such as the one studied may serve to predict learning success to some extent.

For some children there are factors present which impede learning. The child may eventually learn in spite of these factors, but many have not succeeded while they are in the lower elementary grades. By the aid of the Special Education Department 15 unsuccessful pupils were helped to some extent. To discover the etiological factors involved in the lack of academic success involves services and skills not yet available in public school systems in which the students are enrolled.

BOOK REVIEWS

WILLIAM B. MICHAEL, Editor

University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor

Centinela Valley Union High School District

and the

University of Southern California

Coombs' <i>A Theory of Data</i> . RAPHAEL HANSON	621
Butler, Rice, Wagstaff, and Knapp's <i>Quantitative Naturalistic Research</i> . JAMES A. WALSH	623
Miller's <i>Mathematics and Psychology</i> . DAVID M. MESSICK	625
Helmstadter's <i>Principles of Psychological Measurement</i> . KENNETH D. HOPKINS AND B. R. HOPKINS	627
Adkins' <i>Statistics: An Introduction for Students in the Behavioral Sciences</i> . STEPHEN W. BROWN	630
Chauncey and Dobbin's <i>Testing: Its Place in Education Today</i> . PAUL J. HOFFMAN	632
Engelhart's <i>Teleclass Study Guide, Measurement and Evaluation</i> . JOAN J. BJELKE	634
Harris' <i>Children's Drawings as Measures of Intellectual Maturity: A Revision and Extension of the Goodenough Draw-a-Man Test</i> . CAROL HUNTER	636
Harper, Anderson, Christensen, and Hunka's <i>The Cognitive Processes: Readings</i> . SARA W. LUNDSTEEN	638
Manning and DuBois' <i>Correlational Methods in Research on Human Learning</i> . PETER F. MERENDA	640
Valett's <i>The Practice of School Psychology: Professional Problems</i> . MABEL C. PURL	642
Sorenson's <i>Psychology in Education</i> . MABEL E. HAYES	643
Weitz' <i>Behavior Change through Guidance</i> . CHESTER O. MATHEWS	644
Gardner's <i>Development in Early Childhood</i> . EDYTHE MARGOLIN	647
Barshay's <i>Empathy</i> . EDYTHE MARGOLIN	649

A Theory of Data by Clyde H. Coombs. New York: John Wiley and Sons, Inc., 1964. Pp. 585. \$14.95.

This book contains a comprehensive and integrated presentation of the contributions to the analysis of psychological observations made by Coombs and his co-workers. In this analysis Coombs distinguishes three phases of scientific endeavor: (1) going from the universe of potential observations to recorded observations, (2) going from recorded observations to data, (3) going from data to inferential classification of individuals and stimuli. Data, in the terms of Coombs, are recorded observations in which "individuals and stimuli are identified and labeled and the observations are classified in terms of a relation of some kind between individuals and stimuli, or perhaps just between stimuli." Coombs' theory of data is a description and analysis of phase 2 (the creation of data) and phase 3 (the analysis of data).

The chapter organization of the book is based on the 8-fold, triple-dichotomy scheme Coombs uses to categorize the creation of data. The three dichotomies are (1) observations on one set of elements vs. observations on two sets of elements, (2) observations on pairs of single elements vs. observations on pairs of pairs of elements, and (3) observations on an order relation vs. observations on a proximity relation. The justification for this organization is that the data in each octant lend themselves to the same type of analysis.

The focus of attention within the chapters is on methods for the analysis of data. The methods presented for each of the eight types of data are all based on geometric procedures for locating points in unidimensional or multidimensional metric spaces, but where the metric is unknown. The aim of the methods of analysis is the standard one of condensing multitudinous observations into a simple and orderly structure. Coombs catalogues all the methods currently available and discusses in detail the methods originated by himself and his co-workers, particularly J. F. Bennett and W. L. Hays. Thorough and comprehensible treatments of the unidimensional and multidimensional unfolding techniques, non-metric factor analysis, and non-metric multidimensional scaling are presented. The variable of laterality and its ramifications for scaling are discussed.

The heart of Coombs' theory is his "*c, p, q*" model. In this model each stimulus and individual is represented by a point in a r -dimensional space. When individual i is confronting stimulus j on occasion h the individual's location is given by c_{hj} and stimulus' location is given by q_{hj} . The ensuing behavior, which can be classified as either a picking response or an ordering response, is a function of the distance p_{hj} between c_{hj} and q_{hj} . In the general case the parameters are not identifiable because there are many more parameters than items of data. Hence, for applications, important restrictions must be put on the parameters. Coombs does not discuss the various cases of his model, but this reviewer was able to distinguish 18 major cases. Of these, six have identifiable parameters, and two more might be given identifiable parameters with the addition of one restriction on the parameters. Coombs in this book confines himself to the four simplest cases of his general model. For these he assumes (1) the dimensionality of the space is constant for all confrontations of individuals with stimuli, and (2) the positions in the space of the individuals and the stimuli are constant for all confrontations of individuals with stimuli.

Many of the 14 cases heretofore not used could become usable if interval scale or ratio scale values (instead of just the nominal scale values presently represented by the variable j) were given to stimuli based on their physical properties, and if responses were treated quantitatively in similar fashion. Then by quantitatively relating stimuli and responses, empirical numerical laws could be determined. These laws could then be used to obtain additional metric scales for stimuli and individuals. Finally, these laws could be used as an inspiration for and as a check upon assumptions concerning the shifting of individual and stimulus points as they depend upon the pairing of individuals with stimuli in different confrontations. This approach to the richer cases of Coombs' model would introduce into Coombs' theory derived measurement as defined by N. R. Campbell and as advocated by B. F. Ritchie to be essential for the measurement of theoretical constructs and the finding of numerical laws.

The Coombs model is a universal descriptive system because it can be used to describe any set of psychological observations, since a case of the model can always be found which has many functionally independent parameters as there are data. Moreover, it can give a simplified, ordered, and compact description when there are some relationships inherent in the observations.

The reviewer feels that the Coombs model has the potentiality for becoming an explanatory system by the addition of learning and perception postulates which would govern the changes in the individual-stimulus space that are produced by the exposures of stimuli to the individual. With these additions the model would be ex-

planatory because it would entail predictions, and these predictions could be falsified by relevant observations.

The description provided by an application of the Coombs model is usually at the level of an ordering and falls short of a metric description. This happens because Coombs does not wish to postulate a specific metric for the space of points being considered and because he usually does not use metric information when it is given directly in the observations, for he does not trust its scale properties. The strength of the Coombsian descriptive method is that many conclusions about order follow from the sole assumption of an unspecified metric space, and these conclusions can be compared with fairly trustworthy ordinal information given in the observations. The dimensionality of the space and the type of metric can be adjusted until the metric space can encompass the points representing the observations.

This reviewer wonders whether the greater likelihood of truth to be found in the weaker assumptions about model and observations makes up for the considerable loss of precision, power, and efficiency in the process and product of Coombsian data analysis. Coombs' argument on this point, which is merely an expression of personal feeling, is not persuasive (see p. 234). However, even if we grant that more scientific progress can usually be gained by making stronger assumptions than Coombs is willing to make, the Coombsian production of much from little is well worth knowing.

RAPHAEL HANSON
*California State College
at Long Beach*

Quantitative Naturalistic Research by John M. Butler, Laura N. Rice, and Alice K. Wagstaff in collaboration with Sarah Counts Knapp. Englewood Cliffs, N. J.: Prentice-Hall, 1963. Pp. iv + 122. \$4.95.

The purpose of this book is to chart that statistical and analytic wilderness where factor analysis, latent class analysis, and simplex phenomena meet. Under the rubric of analyzing classification systems for naturalistic observation, Butler *et al.* develop a model for finding the characteristic dimensions of a response categorization schema and an index of the internal productivity of the derived system. Along with the model they present the essentials of a computational framework to calculate these characteristic dimensions and to evaluate their statistical index of potential for producing experimentally testable hypotheses.

The technique is closely related to Butler's earlier work on what he has called "successive set analysis" and owes its greatest overall debt to factor analysis. Butler *et al.* argue that a response classification system is first of all a tool for naturalistic observation and not

for hypothesis testing. From their point of view, only classification systems whose dimensionality has been *proven* to be neither trivial nor outlandishly complex are suitable for experimental procedures involving tests of hypotheses. Aside from this, the most radical departure from standard methodology is undoubtedly Butler's refusal to attach any *a priori* values to particular response categories and his insistence on including all categories, even logically exclusive ones, in the initial analysis of a classification system.

Briefly, the model first extracts a number of factors, usually principal axes, from a joint proportions matrix of the binary categories-by-subjects classification matrix. The first principal axis is discarded as a simplex representing what amounts to item difficulties or norms. The next r largest factors are used in direct conjunction with the data vectors themselves to discover the characteristic dimensions of the response classification system and its index of internal productivity. The index is closely associated with the effective rank reduction of the system. As in many factor analytically oriented models, there is one ambiguity about the number of factors to be extracted and about the use of communalities or of some alternative values in the diagonal of the joint proportions matrix.

Three examples are used to illustrate the model. The first, which is an analysis of the Halstead Category Test, must be rated top notch. The book deserves to be read for the sake of this study alone. It not only shows off the model to its best advantage, but also provides an acute insight into a complex problem-solving situation. The other two examples fall far below the level of the first study. They are sparsely described and a little hard to follow. Aside from the undoubted merits of the first example, this entire section would have been much more convincing had conventional factor analyses of the respective classification systems of the examples been included for comparison with Butler's results.

Incidental to the main line of development, Butler *et al.* present a lucid discussion of linear dependencies in binary data matrices and of the properties, particularly with regard to rank, of augmented matrices derivable from a full binary data matrix. Several appendices which are largely superfluous and somewhat repetitious make up the remainder of the volume.

The only major fault of the book is a common one. There is a great disparity in the levels of presentation of the mathematical development and the examples. In all likelihood those who can follow the mathematics will be factor analysts, largely indifferent to the examples; those who might be interested in the examples and in pursuing similar research will typically be oriented toward applied problems and will find the mathematical development incomprehensible. Butler further complicates the situation by indiscriminate

use of several systems of terminology—some going back to Thurstone, some very recent. Because of the disparate levels of presentation this book will prove very difficult for many readers to use and will never reach so wide an audience as its importance and potential usefulness warrant.

JAMES A. WALSH
Montana State University

Mathematics and Psychology by George A. Miller. New York: John Wiley and Sons, Inc., 1964. Pp. x + 295. \$3.45.

This book is a collection of annotated readings selected to illustrate the role played by mathematical thinking in the history of psychology. It is not, however, just another book of readings. It is rather a thoughtful work which outlines some of the scenes in "psychology's long and not always happy affair with mathematics . . ." (p. viii).

Miller distinguishes four major types of mathematical applications in psychology. *Discursive* applications involve the use of mathematical notation and reasoning to extend and elucidate verbal communication. Mathematical language is employed to augment and enrich natural language, not to replace it. This discursive usage is exemplified by selections from Lewin and Herbart.

Next the author discusses *normative* applications of mathematics. This usage attempts to deal with the problems of optimal behavior—what "ought" a person to do. Aspects of the history of normative applications are illustrated by two papers on utility theory by D. Bernoulli and Jevons, a selection on subjective probability by Ramsey, and an introduction to the theory of games by Marschak. Although all of these readings have to do with psychological issues, the authors are not psychologists. This situation reflects the fact that it has been only recently that psychologists have become concerned with the notion of optimal behavior. The development of normative approaches to behavior up to the early 1950's is a chapter in the history of economics.

The largest single class of mathematical applications in psychology is found in the search for *functional* relations among the variables studied in psychological laboratories. The author makes a further distinction between determinate and statistical relations. Determinate functional relations are exemplified with selections from the history of psychophysics (i.e., Fechner, Titchener, and Stevens), learning theory as quantified by Thurstone and Hull, Hecht's theory of the photoreceptor process, and dynamic models of motivation and learning, and social interaction by Simon. The mathematical theories discussed in this class assume a non-probabilistic relation. The discrepancies between data (where data are even considered) and the functional "law" are presumed to be a

result of errors in measurement and other uncontrolled and extraneous factors. Moreover, many of the applications in this class historically have been somewhat unresponsive to problems of measurement. Statistical functional relations, in contrast to determinate functional relations, tend to be more attuned to the necessary interplay between theory and measurement. Probabilistic processes are explicitly incorporated in these models, a fact which renders behavioral variability a concept of theoretical importance. Among the papers in this section are Thurstone's "Law of Comparative Judgment," the Swets, Tanner, and Birdsall paper on "Decision Processes in Perception," and Estes' first article on statistical learning theory.

Structural applications comprise the last major use of mathematical thinking in psychology. The criterial feature of structural mathematics is the emphasis on the analysis of the relations between elements of one or more sets. To illustrate these applications, two selections from the "structure of intelligence" controversy are presented, one by Spearman, the other by Thurstone. In addition there is a paper on the structure of social groups by Festinger, Schachter, and Back, an excerpt from Chomsky's book *Syntactic Structures*, and a selection from Inhelder and Piaget's book on the development of logical thinking in children.

Miller concludes with two papers exemplifying what he calls "quasimathematical" psychology. Both papers are based on the analogy between human behavior and the behavior of a machine. The first is concerned with the human as a seromechanism, whereas the second discusses a computer simulation of a young child going shopping.

For the most part, Miller has done an admirable job in selecting papers which can be appreciated by an intelligent layman or a psychologist with minimal mathematical training. Yet few of the selections would be boring to the technical reader. This fact alone is quite an accomplishment. The major achievement of the book, however, is the continuity which results from Miller's desire to recount systematically the story of psychology's flirtation with mathematics. In doing this, the author has written nearly one page of commentary for every two pages of readings. In these comments he provides transitions from one selection to the next and often fills in the gaps with descriptions of intervening ideas and research. The result is that, for the most part, a coherent and readable story develops.

Occasionally, the tale becomes a bit sporadic and discontinuous, with loose ends here and there. This fact is more a reflection of the history to be told, however, than of Miller's way of telling it. Psychology has often adopted mathematical approaches which have originated in other disciplines. Many of these approaches have been

adopted only recently moreover, as is the case, for example, with game theory, graph theory, Markov processes, information theory, and computer simulations. Psychologists have just begun using these tools within the past ten or fifteen years. There is little to document historically other than the fact that the introduction has been made and an intimate relationship seems to be emerging.

There are two facets of psychology's fling with mathematics which are unfortunately not outlined in this book. First, there is the story of the psychologists' use and misuse of statistical techniques in data analysis. This tale is quite interesting and often enlightening. Second there is the development of psychological scaling, an important and relatively mature application of mathematics to psychological problems. Although the author has good reasons for not wanting to broach these topics in a book of this size, the history would have been much more complete had he done so.

The book should make a significant contribution to the teaching of courses in the history of psychology and in introductory mathematical psychology. In the former, it should serve as a source of ideas for students with a quantitative penchant and as a source for original readings to supplement a standard text. In courses in mathematical psychology, it will be of considerable value in helping to provide an historical context for contemporary problems and in documenting the wide diversity of psychological problems to which a mathematical approach has been made.

DAVID M. MESSICK

University of California, Santa Barbara

Principles of Psychological Measurement by G. C. Helmstadter.

New York: Appleton-Century-Crofts, 1964. Pp. ix + 248.

Helmstadter has achieved well his objective to present a text which "... concentrates on the underlying principles of testing rather than on the instruments themselves." Although designed for an initial course in testing, the book assumes an elementary knowledge of statistics. The reviewers would have felt more comfortable if the meaning of certain technical terms had been briefly redefined and reviewed as they were introduced, e.g., normalized (p. 42), *MS* (p. 73), covariance (p. 64), and binomial distribution (p. 176) among others; the initial pupil inoculation with statistics often does not "take." The somewhat greater depth in the presentation of basic concepts, however, was received as both refreshing and salutary. When the senior reviewer was asked to evaluate this work, he had just concluded the teaching of a course in which he had selected it as one of two principal texts. It is felt that this experience has sensitized the reviewer to certain factors relating to the book's "teachableness," via student questions and comments.

In ten brief chapters of 232 pages, the fundamental topics of standardization, reliability, validity, test development and analysis, and multiple measurement are clearly and concisely presented. There are also introductory and concluding "overview" chapters; the former, in being uneven in depth and content, appeared to be less than adequate. Rather than giving the reader the intended "Gestalt," it seems to present "too much, too soon," at least concerning the topics of construct validity and expectancy. The final overview chapter is in sharp contrast; it more than compensated for the deficiencies of the initial chapter. The chapter gives an unusually succinct and integrated summary of almost every basic concept introduced in the previous chapters—an achievement which will benefit the learner greatly.

The book's most severe deficiencies are found in the chapter on standardization and norms. It contains serious omissions of several basic topics, as well as incomplete coverage of others. Approximately one-half of the brief 22 pages is not *directly* related to standardization. It would seem to belong elsewhere. For example, variable and constant errors, as well as expectancy tables, could more profitably be placed in the chapters on reliability and validity, respectively. *T*-scores are unfortunately presented as non-normalized. There is no discussion of the much-used stanine or deviation IQ scale, nor of the problem of variable sigmas with ratio IQ's. No mention is made of the various methods of determining and defining grade placement values; for example, modal norms are not treated. Perhaps a more serious omission is that of the Buros' *Mental Measurements Yearbooks* and *Tests in Print*, neither appears in the text or references. Certain imprecise and misleading statements also appear. For example, "In testing, this problem of interpretive errors is taken care of through a process called *standardization*" (p. 36), and "... validity requires some minimally satisfactory degree of objectivity, reliability, and *standardization*" (p. 40). [All italics in quotations are those of the reviewers.]

The chapter on reliability is excellent. A lucid presentation of true and error variance, as well as of certain of the more common methods of estimating reliability, appears. The reviewers were pleased to note the inclusion of Hoyt's method, relating reliability theory with analysis of variance, but were disappointed to find no mention of the commonly-used Kuder-Richardson (KR) Formula 21, nor any discussion of lasting, temporary, general, and specific types of variance. Unfortunately, certain inexplicit and inaccurate expressions appear: (1) "Interdependent or nearly identical items will *reduce* reliability," (p. 81), (whereas the converse is actually the case); (2) The "index of reliability" (p. 225) was used imprecisely as being synonymous with reliability coefficient; (3) It is stated that reliability estimates treat "... *any and all changes in*

score from one measurement in time to a second such measure as being error" (p. 65), (which is true, of course, only if there is variability in change, not if there is change *per se*); and (4) "... the parallel form reliability is preferred *because* it provides a conservative estimate..." (p. 225).

The topic of validity receives extended attention (three chapters, 66 pages). The chapter on construct validity is perhaps the best treatment in any available text; the concepts of Cronbach and Meehl plus the multitrait-multimethod approach of Campbell and Fiske are clearly and completely presented. The reviewers are of the opinion, however, that the discussion of factorial validity and of the validity of various theories and measurement approaches to the construct "intelligence," which appear in the chapter on content validity, fall more naturally in the domain of construct validity—a fact which increases the difficulty in differentiating between these two types of validity. An additional undesirable organizational feature, which will create ambiguity for some students, results from the inclusion of a 12-page section, "Special Problems in Interpreting Test Results," only marginally related to the chapter theme, "Construct Validity."

The presentation of the standard error of estimate will be confusing to some—a confusion resulting from an unnecessarily frequent shift between $\sigma_{y,x}$ and $\sigma^2_{y,x}$. In addition, no mention is made of the requirement of homoscedasticity, if a single value is to be appropriate for all score levels. Terminology is sometimes unconventional: the coefficient of determination appears only as the "relative reduction in error," and the success ratio following the use of a test is termed, "validity rate."

Certain perplexing statements pertaining to validity also appear: (1) "When used for research purposes, tests must usually be considered as *completely valid* measures of certain human characteristics" (p. 9); (2) "If either the criterion or the test or both variables are not measurable on a *ratio scale*, then some other form of correlation (than Pearson product-moment) such as rank order, biserial, tetrachoric, or the correlation ratio may be used" (p. 112); (3) "Although there may be some exceptions, it has generally been found that when a test is carefully constructed and appropriate item evaluation and selection procedures employed, a *completely satisfactory degree of validity and reliability* can be achieved with a relatively short test" (p. 174). The definitions given for non-language and speed tests were incomplete or inaccurate: "Non-language tests are those which, in contrast, to most tests use *no* written or spoken word in either the directions or the test" (p. 13); "A test is referred to as a *pure speed test* if it is such that *everyone who reaches an item gets it correct*" (p. 79). [All italics in preceding quotations are those of the reviewers.] Certainly an untimed test

in which *all* examinees get *all* items correct is not a speed test, even though it fits Helmstadter's definition.

There is an excellent treatment of item analysis and item statistics. The reviewers wished that the author had made the point that his suggestions were geared specifically to standardized testing. As it stands the naive student is apt to feel that an item analysis on classroom tests is not worthwhile, since the recommendations of the chapter are rarely achieved—for example, 400 subjects and several forms of the test differing in item order. It was also regrettable that only one measure of item reliability (item-test correlation) and only one index of item validity (item-criterion correlation) were given, since each one presented is not very commonly used, in test studies, and is more time-consuming than are most other indicators. Its purported advantage (being able to estimate test validity by dividing the average item validity by the average item reliability) is not an adequate compensation for the confounding of item reliability (discrimination) and difficulty. That is, items can differ markedly in "item validity," even though they correlate equally with the criterion. Furthermore, the "advantage" is more apparent than real, since item validity data are often absent because of the unavailability of an external criterion.

Several printer and/or editing errors appear: X for X_0 (p. 115), diagnosis for diagonals (p. 142), Figure 25 for 24 (p. 150), 67 for 53 (p. 167), and an omitted "not" from a sentence (p. 152). In fact the major weaknesses of the book could have been avoided with careful organizational and editorial assistance.

Although the reviewers have felt it necessary to bring to the reader's attention specific weaknesses of the work, they are most hopeful that its many merits and strengths are not consequently overshadowed. When a relative yardstick is used, the book is strong; the senior reviewer is continuing its use as one of the two principal texts for a course, "Fundamentals of Measurement," designed primarily for the training of counselors and school psychologists. An instructor can easily clarify the points needing attention noted in the review.

KENNETH D. HOPKINS
University of Colorado
AND B. R. HOPKINS
Biola College

Statistics: An Introduction for Students in the Behavioral Sciences
by Dorothy C. Adkins. Columbus, Ohio: Charles E. Merrill
Books, Inc., 1964. (Hardcover \$8.95, paperbound, \$6.95.)

This new text was designed to be used in a one semester introductory statistics course for students in the social sciences and education. For the most part, the topics covered and the order of

presentation are very similar to the introductory chapters of the more familiar texts in this area. Exceptions to this similarity may be seen in that (1) a separate chapter is devoted to the special correlation techniques and (2) analysis of variance is not discussed.

One of the biggest assets of this book is Professor Adkins' superb style of writing. In the early chapters the student is led through an excellently written, informal, and intuitive review of elementary algebra and introductory descriptive statistics. Following these informal beginnings the chapters gradually change toward presenting step-by-step derivations of formulas and toward defining terms in a precise and structured manner. Rigor and formality are constantly being increased, but never in a threatening manner. This gradual building up of sophistication, as well as the excellent review of algebra, allows almost any student, even those with a meager background in mathematics, to master the concepts and formulas of elementary statistics.

Another very favorable aspect of this volume may be seen in the use of a unique self-testing feature. At key points within the text and at the end of each major topic the student is confronted with a multiple choice question concerning the material just covered. After deducing an answer the student moistens a point corresponding to one of the alternatives, and a color change provides immediate feedback on the correctness of his response. In addition to the motivation added by the novelty of this device, the immediate feedback also offers the student an opportunity to go back and to master the material just read before he proceeds to other topics.

As is the case with most texts, some topics are necessarily omitted or given only token coverage. One of these omissions—analysis of variance—has previously been mentioned. In addition, it should be noted that there is almost a complete lack of the traditional statistical exercises, and only three tables are included in the appendix (z , t , and χ^2). Furthermore, the reviewer believes that the material concerning the testing of null hypotheses could be greatly expanded and elaborated. Relative to this point it is worth noting that the themes of inferential statistics and confidence intervals are excellently discussed throughout all phases of this book. However, the last chapter titled "Testing Statistical Hypotheses" does not seem to integrate all the issues, nor to offer much on the formulation of null hypotheses—all of the important points are mentioned, but they are only briefly developed.

Although there are a few other minor weaknesses (such as a somewhat difficult-to-follow illustration concerning estimation of a Pearson r from a scatter plot or the omission of both ABACS and approximation formulas for the special correlation techniques), the reviewer believes that this volume will be very well accepted by most students and teachers. Considering all factors, however, it is

felt that this book should not be used as the main text in an upper division statistics course, but rather should be included as a required supplement to one of the larger, more comprehensive texts. In this way the student will be given the advantage of Adkins' lucid and easily followed writing style, while he is using the other text for its exercises, tables, and more formalized discussions.

STEPHEN W. BROWN

University of Southern California

Testing: Its Place in Education Today by Henry Chauncey and John E. Dobbin. New York: Harper and Row, 1964. Pp. 223.

The most unimpressive feature of this book is the design of its jacket—an unimaginative blocking of dingy brown and black, with white and red lettering that is neither easily readable nor attention-getting. The college teacher and assessment researcher will disagree; they will be inclined to judge the contents as at least equally unimpressive, for the book represents an over-simplified and abbreviated collection of generalizations about tests and testing in schools.

The book claims to be "... an authoritative and balanced discussion of what tests, as partners of teaching, can and cannot do—with advice to both parents and teachers." It is. And what is more, the message of the book has to do with erasing from the mind of the public a great deal of negativism and hostility about tests and testing—attitudes which have originated from abuse and misuse of tests and which have festered as a result of professional aloofness of psychologists vis-à-vis the public. These negative attitudes have been nurtured through popularized and often exaggerated or erroneous accounts of the fallacies and weaknesses of objective tests. In these respects, Chauncey and Dobbin do more than erase. They have provided the type of discussion any parent or high school child can follow—factual, authoritative, and well calculated to supplant warped notions about testing with favorable ones. Since these positive attributes of the book have been well expounded by D. P. Campbell (see *Contemporary Psychology*, Dec., 1964), this review will be a somewhat critical appraisal of particular features of the book.

The book is short, and the treatment brief. The first five chapters (81 pages) provide a thumbnail sketch of the history of testing, aptitude and achievement measurement, norms, reliability, validity, sampling of content, the interpretation of test scores, and a discussion of the relative merits of essay vs. objective tests. This discussion is followed by a chapter on tests as tools in teaching, a chapter on tests in selection and admission, one on tests in guidance, and a final chapter for the student, filled with helpful hints on how to do well on an examination.

The writing style is deliberately simple, and some sections are so "homey" as to be indistinguishable from a Dick and Jane novel; for

example, "The admissions officer then writes a warm letter to Fred and—when the letter arrives, Fred and his parents are so pleased that they go out for dinner to celebrate" (p. 130). Some may find disturbing the fact that Chauncey and Dobbin do an excellent job of representing the Educational Testing Service (ETS) as a leader in its field in general, and as a publisher of the *Sequential Tests of Educational Progress (STEP)* in particular. This reviewer welcomed the rather detailed presentation of the *STEP* tests as illustrative of up-to-date methodology put to work.

What this reviewer found disturbing was the deliberate obliteration of the classical distinction between aptitude and achievement. For Chauncey and Dobbin, aptitudes are meaningless concepts, and all intellectual tests are "work-samples," which measure present knowledge and skills. This theme, which is recurrent throughout the book, tragically reflects what is probably the dominant position in education circles today. It is, in fact, a denial of the importance of intellectual aptitudes, and it leads the authors to adopt the "work-sample" concept of intelligence. What is more, it leads them to a purely cultural definition. For Chauncey and Dobbin, intelligence is the measurement of what "learned" people do, and since people do different things in different cultures, there is little need to talk about the ability to function at an abstract or symbolic level. For this reviewer, cultural definitions of intelligence are a menace to psychology, as well as to education. Consequently, the reviewer cannot appreciate the conclusion reached by the authors that tests which predict the success of Australian aborigines in hunting and fishing are "intelligence tests" (page 22). The absence of a frank discussion of classical intellectual aptitudes, largely genetically based, is not at all helpful to teachers, parents, or pupils insofar as effective guidance and decision-making are concerned.

A few minor criticisms may also be made. On page 62 we find that 40 to 50 objective questions are the usual minima for reliable measurement, a test length which would hardly suffice for the teacher who habitually uses true-false examinations. The discussion of score interpretations on the *STEP* (pages 52-53) confuses the error of measurement with the standard error of a difference. Although in Chapter 5 the discussion of objective vs. essay tests is clear, it is colored by the same attitudes about intellectual attainment which run rampant throughout American educational institutions today. The reviewer is referring to the traditional treatment of education as a passive process of "soaking up" verbal and other symbolic knowledge, from which it follows that what one wants to measure is the retention of symbolic knowledge. Of course this type of retention can be best estimated through use of objective tests, which usually require only the ability to recognize. However, the real goals of education can be construed as involving much more

than a recognition of verbal relationships. These goals can include the active searching, ordering, deducing, analysing, and even reconstructing of material into forms which can be defined objectively only with great difficulty, if at all. The essay examination, like the oral examination, the class discussion, the painting, and the entry for the Science Fair are obviously the best and the most direct work-samples of these educational goals. Teachers have traditionally shown some reluctance to include essay tests in pupil evaluations, perhaps *because* in their subjectivity they expose the teacher's judgment to criticism. Objective tests provide the escape. Although in many instances objective tests can suffice, there are still important occasions in which they cannot. Despite the fact that the authors recognize this limitation, they treat these aims of education involving problem solving and higher level psychological processes only as rare exceptions.

Chapter 6 provides a good discussion of practical problems in the classroom as well as a treatise on the need to furnish a level of instruction commensurate with present achievement. However, the bias against the apparent existence of aptitude prevents the authors from clarifying the problem of readiness. For Chauncey and Dobbin, tests indicate to the teacher differences among students in present level. Thereby a choice is allowed in the selection of curricular materials. However, since tests most often indicate differences in *rate* of learning as well as differences in level, it hardly seems appropriate to plan the same instructional unit for the ten year old as for the eight year old because both have attained a nine-year level of proficiency.

Nevertheless, these weaknesses do not detract from the purpose and message of the book, insofar as the public is concerned. It is the reviewer's hope that Harper and Row will design a new dust jacket for the book. Such an aesthetic improvement will help this volume to find its way into a large number of households.

PAUL J. HOFFMAN
Oregon Research Institute
Eugene, Oregon

Teleclass Study Guide, Measurement and Evaluation by Max D. Engelhart. Chicago: Chicago Public Schools, 1963. Pp. iii + 110. \$1.50.

Originally compiled as a teaching aid for a tests and measurement class produced by WTTV, Channel 11 Chicago, this study guide contains basic tools which the student may use both in preparing for and reviewing of lecture material covered in a measurement class. This instructional aid was designed to be used in conjunction with two texts: (1) the basic text required of all students: *Introduction*

to *Educational Measurement* by Victor H. Noll, and (2) a supplementary text, *Essentials of Psychological Testing*, Second Edition by Lee J. Cronbach. Included in the "Course Outline," pp. 8-10 are reading assignments in Noll paralleling the lecture topics. In addition, numerous selected references for in depth reading follow each "supplementary reading" section.

This soft bound manual serves two purposes. (1) Initially it delineates the scope and requirements of the course and provides the student with a discussion of the instructional objectives and course outline. (2) In addition, supplementary readings are included in the following areas: (I) "Test Reliability and Validity," (II) "Instructional Objectives," (III) "Short Answer and Essay Exercises," (IV) "Suggestions on the Form of Objective Examination Exercises" by Macklin Thomas, (V) "Item Analysis," (VI) "Interpretation of Test Scores and of Differences between Scores," and (VII) "Using Test Data in Classroom Experimentation."

For the most part, the supplementary readings are designed to introduce the theory and concepts behind the major topics. This guide does not attempt to provide the student rigorous enough coverage to be the only source of reading and study. However, it does lay the framework through which the student may "actively" participate in even a one-way lecture over the television media. One of the major exceptions to the conceptual coverage is in the final section on the use of test data. The author presents practical applications of matched group comparisons, analysis of variance, and analysis of covariance. In this chapter actual step-by-step procedures are outlined. Throughout the readings, adequate formulas, drawings, and tables are included to facilitate presentation of lecture material without requiring the student to reproduce intricate charts and graphs shown over a television screen. This reviewer feels that providing students with such helpful teaching aids also has merit for the regular classroom instructor, who in otherwise being forced to fill the board with numerous charts, graphs, formulas, and problems throughout the lecture period, thereby draws from valuable lecture and discussion time.

This reviewer feels that the addition of a few study questions and problems following each reading section would enhance the value of this publication. In spite of this relatively minor limitation, the reviewer feels that this study guide could be used to good advantage in a regular classroom situation. Properly utilized, this instructional aid could serve as preliminary preparation for students for lecture and discussion sessions. And, the efficiency of having formulas and problems in the hands of each student could greatly encourage active student participation in the classroom.

JOAN J. BJELKE

University of Southern California

Children's Drawings as Measures of Intellectual Maturity: A Revision and Extension of the Goodenough Draw-a-Man Test by Dale B. Harris. New York: Harcourt, Brace and World, Inc., 1963. Pp. vii + 367. \$8.95.

In this book the author sought to revise the Goodenough *Draw-a-Man* test (issued in 1926) and to finish uncompleted aspects of Goodenough's research. The research included the following efforts.

- (1) An unsuccessful attempt was made to extend the Goodenough scale to include adolescent years. (It was found impossible, as most children in an overwhelmingly visual culture are so critical of their ability that they give up drawing. Educators need to take note and to make provisions.)
- (2) An alternate form to the *Draw-a-Man* scale was successfully devised through the development of an analogous point scale for the "Draw-a-Woman."
- (3) A drawing of the "self" was also included as a possible third form in the author's attempt to discover a more valid projective device for the study of affect, interest, and self-concept than an impersonal figure. The author specifically requests that the child draw his "self" and provides a guide for analysis. However no empirical tests have yet been reported. Although a warning of conservatism is set forth, few suggestions of a constructive nature are offered.
- (4) A "Quality" scale for the man and woman drawings was constructed and standardized for quick approximation to point scores. The author maintains that a quality scale does discriminate conceptual development as adequately as does a point scale for children from five to nine years of age. The reliability estimate was judged to be about .90.

For the purposes of the school psychologist the "Quality" scale shows promise. One of the prime roles of schools today is prevention of school maladjustment. The Goodenough-Harris Drawing Test should be easily integrated with other screening tests into a battery which can be used to select children who should receive more detailed attention. Children can be quickly arranged in order of intellectual maturity in kindergarten and first grades when no group mental maturity test data are available.

Since the old scale of 51 scoring items was oppressive in the eyes of many primary teachers, the new scale of 73 points may be even more difficult to accept than the older one. With the degree of correlation between the Goodenough *Draw-a-Man* test and individual intelligence tests as modest as it is, why would this qualitative scale not serve for screening? The reviewer recommends further studies in this area.

After the procedures involved in the development of Goodenough's *Draw-a-Man* test (1926) were reviewed, the present revision was explained. The chapter on methodology of the revised scale contained specific data on the validation of items for the scales as well as information concerning special scoring and scaling problems. The change from mental age to standard score and percentiles seems advisable.

Reliability and validity of the 1926 and revised scales are discussed at great length. Although estimates of reliability appear to be satisfactory, validity coefficients of Goodenough's test with individual intelligence tests are uniformly positive, but range from a very modest value of .20 to quite substantial magnitudes of .70 or .80. A table on pages 96 and 97 summarizes correlations between Goodenough's 1926 instrument and other psychological tests. Further validity data will need to be reported on the revision.

The revised scales—*Draw-a-Man (DAM)* and *Draw-a-Woman (DAW)*—appear to be adequately standardized. Representative geographic areas and occupational distribution were used. Socio-economic status was assumed from occupation. Children of equal month of birth were selected. Frequency of sex membership was equal. Ethnic groups were not discussed.

Definite sex differences were noted and discussed in detail. Although girls were favored on both scales, the difference was even greater on the *DAW* than on the *DAM*. Instead of minimizing these differences as most psychometric instruments do by excluding items favoring one sex, the author feels that these differences should be taken into account. His scales provide separate tables for boys and girls.

In discussing and reporting cultural influences in drawings, the author concluded that although the test may be unsuited to comparing children *across* cultures, it still may rank children *within* a culture according to intellectual maturity.

In addition, the author presented a comprehensive survey of empirical and theoretical literature on children's drawings. The author offered some constructive criticism of the research methodology as well as indicated studies which might have been designed.

School psychologists should find the Goodenough-Harris *Draw-ing Test* useful in screening primary children. Clinical psychologists should find the information provocative. One may hope that the ingenuity of psychologists is taxed in formulating experiments based on theories in order to clarify some of the issues which the author discussed.

CAROL HUNTER

*Torrance Unified School District
Torrance, California*

The Cognitive Processes: Readings by Robert J. C. Harper, Charles C. Anderson, Clifford M. Christensen, and Steven M. Hunka (Editors). Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1964. Pp. xii + 717. \$8.85.

The volume *The Cognitive Processes: Readings* represents a collection of contributions from the past decade to the understanding of various systematic approaches to the study of the thought process. The 47 contributions present ideas that are simple but with both articulation and implementation that are difficult.

Consisting of 47 chapters the volume is well organized about six major units: "Motivation" (Part I), "Neo-Behavioristic Approaches to Cognition" (Part II), "Information Processing Approach to Cognition" (Part III), "Computer Model" (Part IV), "Cognition, Motivation and Personality" (Part V), and "Cognition in Children and Cognitive Development" (Part VI). The several theoretical positions presented give insight concerning (a) recurring themes, (b) shifts in emphasis, and (c) incompatibilities. Similarities are also revealed which may stimulate further search for points of contact between informational and behavioral theories.

If one directs attention toward the implications of this book for mental measurement, intelligence testing, and individual differences—the concerns of *Educational and Psychological Measurement*—articles from Bruner are perhaps the major contribution. Found in Part III, his work represents an emphasis on stimulus inputs and on the structuring of information into definite forms or models. Central to higher-level mental processes is the development of a symbolic model, consisting of a structure, and a system of categories (a generic coding system) which represents environmental information in an economical manner. Imposition of this structure can also transform input in such a way that new information is generated. In his article on perceptual readiness, Bruner defines perception as an "act of categorizing."

From reading Bruner's articles it is possible to infer that testing should be concerned with the pupil's grasp of the structure of the subject matter which provides a model or coding system. Or testing should be concerned with the ability of the pupil to generate new information (pp. 293-311). In effect, the process of thinking rather than the product should be the focus of the testing. Or it may be found that there is an interesting correspondence between the psychological determinants of test performance and Bruner's coding systems.

In Part IV, Newell, Shaw, and Simon show that the behavior of the computer in terms of intermediate and end products appears in some ways comparable to human behavior in a similar situation. There may be similarities between the strategies required in com-

puter simulation and those postulated by Bruner as being required in concept formation tasks. But in one's searching for implications for mental measurement, it is well to remember that computer simulation is only an approximation to the understanding of cortical activity. Human variables exist to present problems for the computer model—for example, (a) imperfect reception and transmission, (b) the unpredictable variation of subjective probabilities, (c) the acceptance of verbal information not fully comprehended, and (d) the reading off of information beyond that conveyed by the symbols observed.

For those concerned with individual differences, perhaps the major contribution is made indirectly by the Kendlers (Part VI) in the article "Vertical and Horizontal Processes in Problem-Solving." Kendler and Kendler utilize the construct of the response-produced cue as a mediator between external stimulus and external response in problem-solving. The experimental task or test involved cups or containers differing in size and brightness with regard to one dimension. To phrase it in a direct but somewhat oversimplified manner, they found that slow children and young pre-school children reacted to their tasks using reversal shifts more like rats, whereas older and brighter children tended to react in terms of their mediation theory. Moreover, the Kendlers point out the importance of the variable, verbal learning, in the task. Verbal learning as found in the classroom is neglected in this volume as are the mental processes of critical thinking and creative thinking.

Unfortunately, recent attempts to test the higher mental processes of thinking in classroom experimentation by Hilda Taba of San Francisco State College and Richard Suchman of the University of Illinois were made too late to appear in this volume. For persons concerned with implications for educational measurement, the present reviewer would like to nominate these two studies as the outstanding research in cognitive processes of the past ten years. Classroom experimentation is ignored in the present volume. Perhaps this neglect reflects a point of view.

There has been emphasis in the past on applying the findings from the laboratory to the classroom. A reverse flow of ideas might also be useful. Studies of classroom experience with regard to thinking may contribute to the focusing of hypotheses of cognitive development for the laboratory and may possibly influence methodology in such a way that one progresses beyond the "little cup—big cup" stage referred to earlier. The child's actual problem in this kind of task may be to please the research worker or to finish as soon as possible. The problem task to a large extent must be the child's own in order to call forth a valid response. Instead of approaching the big and complex solely through the little and the simple, perhaps it is time to confront the big and complex directly, or at least to at-

tack the problem of cognitive development simultaneously in the classroom and in the laboratory.

In summary, the authors of this book would agree with Guilford in Part IV that although the development of intelligence testing has received most of its impetus from factor analytic techniques, "You cannot get out of an analysis what you do *not* put into it" (p. 367). Accordingly, if it can be argued that a knowledge of the different attacks that can be made on the nature of cognition and cognitive development will help to solve the problem raised by Guilford, then this book may be of some value to those dealing with mental measurement. There may, for a final example, be an interesting correspondence between the psychological determinants of test performance and Piaget's conception of concrete and formal operations (p. 317ff). New avenues for future cross-fertilization between intelligence testing and cognitive theory may be opened. Besides the consideration of the above content by theorists in testing, in reverse direction, it would be of value to theorists in cognition to have the measurement tools from the test theorists for support of longitudinal, comparative, and programatic investigations of thinking.

SARA W. LUNDSTEEN
University of California,
Santa Barbara

Correlational Methods in Research on Human Learning by Winton H. Manning and Philip H. DuBois. Perceptual and Motor Skills, Monograph Supplement 3-V15, Missoula, Montana: Southern Universities Press, xv (1962), 287-321. \$2.00.

In this interesting and informative monograph on correlational methods the authors address themselves to the question of how correlation techniques can efficiently be applied to the measurement and analysis of change in human learning. Several research approaches are discussed from a theoretical standpoint, and a number of actual research studies are reviewed to illustrate the practical application of the correlational methods (including factor analysis) proposed by the authors to measure the relationships among learner characteristics as well as later changes in proficiency and skill after the learning sessions. The general topics of the training research studies discussed are the following: (1) The Prediction of Gain in Educational Psychology; (2) Prediction of Retention of Technical Knowledge; (3) A General Factor in Psychomotor Learning; and (4) A General Factor in Learning of Electronics.

In their theoretical discussion, the authors argue that the more appropriate criterion variables in studies investigating the correlates of improvement associated with learning treatments are measures of residual gain rather than of the conventional crude gains. They then proceed to reanalyze some previously published data to illustrate

the relative efficiency of residual gain measures, and they go on to report on four studies which were specifically designed to incorporate within their design the employment of residual retention measures as the criterion of change. The residual value in these studies is that portion of the criterion variance in these studies which is not predicted from an earlier measure. The validity coefficients, therefore, are part correlations, resulting from variance removal by partial reduction methods. The results of these studies have led the authors to conclude that "... the use of residual gain as a criterion for the validation of selection tests would serve to facilitate selection procedures oriented toward criteria of training ability or educability, rather than to achievement at a particular point in learning."

The authors' thesis appears to be upheld by the rigorous theoretical formulations by the authors which precede their numerical illustrations which utilize applicable and extensive, although rather dated, information. However, there are unwarranted liberties and casual oversights which the authors commit with their data which require some comment. Perhaps in their eagerness to include in their analysis as many independent variables as possible, phi coefficients based on true dichotomies are treated as correlation coefficients, and repeated reference is made to these as measures of relationship rather than association. Also, in their studies utilizing naval school enlisted men as subjects, the authors casually state that the restriction in range present in some of their principal predictors—the Navy Basic Battery Tests, "... is probably a result of the prior selection of trainees for the Aviation Electronics Technician School." This should not have been a contingent statement. Numerous published studies by the Navy, some by this reviewer, present unequivocal evidence of the actual selective nature of these test scores among Navy Class "A" and Class "B" School students. It is this reviewer's opinion that Manning and DuBois not only should have acknowledged this fact, but also should have corrected their correlations by multivariate techniques as do the Navy research psychologists, for the multiple restriction in range which is known to exist within these samples. Had they done so, the results would likely have been different—ones which would have further reinforced and supported their arguments.

Aside from the above criticisms, this reviewer has found the monograph to be interesting and lucid. It should be studied by psychologists conducting research on human learning, for the sound methods described by the authors appear to offer promise of improved and more encouraging results in analyzing changes in proficiency and skill gained through formal training procedures.

PETER F. MERENDA

University of Rhode Island

The Practice of School Psychology: Professional Problems by Robert E. Valett. New York: John Wiley & Sons, Inc., 1963. Pp. x + 339. \$7.50.

This book is one of the few written which have concerned themselves solely with the profession of school psychology. All have appeared within the last few years. Conference reports and journal articles have long reflected the concern of school psychologists regarding who they are and what they do. This book attempts to answer these questions for school psychologists themselves and for "... citizens and professional persons who are hearing more and more about this relatively new and rapidly growing profession." The author's answers consist of descriptions of those things which school psychologists actually do, and the problems they face in so doing.

The reader experiences a somewhat overwhelming picture of the functions of a school psychologist, as detailed descriptions of the technical problems as well as those of professional relationships are presented. In attempting to present the complete picture, the author falls victim to the necessity of making statement after statement without being able to develop his ideas. One is impressed by the fact that the author knows a great deal about school psychologists and their operations. It would be difficult to believe that anyone currently so employed could not find himself and his problems outlined within the pages of this book.

The problem situations presented throughout the volume are varied, and are frequently not easily answered. However, they represent situations which are actually encountered. The case material used also is adapted from actual practice. It runs the typical gamut from commonplace to bizarre.

In the opinion of the reviewer, the chapter on psychological theory is less effective than are those chapters concerned with practice. For the reader who lacks a background of theory, the sections on mental organization, concepts of self, the development of self-awareness, personal integration, and learning are inadequate in depth to provide a conceptual framework. For the reader with theoretical sophistication, the sections add little. The reviewer does not disagree with the author on the importance of theories for the school psychologist and with their implications for practice. In all fairness, it must be noted that the author stated that his purpose was not to present in detail, but to integrate the theories as a meaningful background and introduction for subsequent material. This purpose was not entirely successful in accomplishment.

Although the book in the final section discusses broader aspects of professional practice, its main concern is with the practical problems of the school psychologist whose role is primarily case study, ex-

amination, and consultation. To this extent, the book should be particularly useful for stimulating discussions in graduate courses on in-service training programs.

MABEL C. PURL
Riverside City Schools
Riverside, California

Psychology in Education (4th Edition) by Herbert Sorenson. New York: McGraw-Hill Book Company, 1964. Pp. 555.

The present volume is an overview of the significant aspects of child and adolescent growth and development; the techniques used in the evaluation of intelligence, personality, and achievement; the application of the fundamental principles of integrated learning; and the educational implications of personal and social development of the learner.

This comprehensive coverage, which summarizes the numerous and relevant factual and theoretical studies that have been made of many aspects of educational psychology, is intended to bring together such new information in the field as seems germane with the best of what was previously known and thought. Those references cited by no means exhaust the relevant literature, but do include a large proportion of the more adequately conducted studies in the field.

Although the organization of the book resembles that of most widely used texts, the mode of approach employed in the presentation of the data accentuates the readability of the text. The author makes a conscious effort to support modern theory with experimental findings through their close integration. The objective is to present substantial content which is validated in every possible particular.

A tenet of this text is that the classroom is only one part of the proper sphere of educational psychology. The problems of educating the individual are treated in terms of the larger context—the home, the school, and the community.

The main areas under discussion are divided into four sections, titled, "Human Development," "Individual Differences," "Mental Health and Behavior," and "Learning." In addition to its emphasis on the main areas under consideration, significant sub-chapters are devoted to the discussion of human needs and motives, developmental tasks, socioeconomic class and its impact on the school, and the personal and professional development of teachers.

This book is enhanced by its clear, simple style and by summaries accompanying each chapter. The new edition is designed to be interesting and teachable. At strategic places in each chapter are "Exhibits," and "Consider the Data" and "Reflect and Review"

sections which encourage the student to enlarge his knowledge and to clarify his thinking.

The text can be used with profit, supplementing the basic textbook in a psychology course.

MABEL E. HAYES

University of Southern California

Behavior Change through Guidance by Henry Weitz. New York: John Wiley & Sons, 1964. Pp. xiv + 225. \$6.50.

The title of this book reflects its central theme. However, if one expects to find specific answers to a "How do you do it?" question, he will be disappointed. The main purpose is to focus attention on the social, global, nature of the behavior of the unique individual and to describe a conceptual framework for understanding and changing this behavior. The emphasis is on principles rather than methodology. In words from the author's preface:

Here we have tried to examine the structure of idiosyncratic human behavior, to see the interactions between a unique biographical history and a unique configuration of circumstances, and to relate this behavioral interaction to that crucial human activity of collaborative problem solving that we have come to call guidance. This is no fully formed theory of behavior. . . . It is, however, a sincere attempt to indicate trails leading to vantage points from which behavior may be viewed and to suggest particularly interesting vistas that may begin to give a new perspective to guidance. (p. ix.)

One of the main contributions Weitz has made is the difference between objective and symbolic reality and the necessity of being able to translate behavior to symbolism to study it and symbolism to behavior to change it. Observation, recording, and analysis of the continuous flow of behavior is possible only through translating it into "the symbolic reality of our minds and our language" (p. 6). In this symbolic form structural planning takes place by client and counselor working cooperatively. This plan must then be activated by translating it back from symbolic to objective reality. Furthermore, no change beyond solving the immediate problem will take place unless the client learns (and it is a learning process subject to all the principles of learning) to generalize this experience of problem solving to other similar behavior.

The first three chapters deal with "The Structure of Behavior," "Components of the Behavior Product," and "Modification of Behavior." This idea that all behavior is the product of one's past (his reactional biography) and the environmental context, as these interact to form the stimulus function, is applicable to all learning whether it be called instruction or guidance. Weitz holds that there

are significant differences between both the goals and the methods of these two functions of education—instruction and guidance.

The remainder of the book deals with the guidance function in the light of the theoretical formulations developed in these early chapters. There are chapters on "The Guidance Function of Education," "Problem Identification," "Measurement in Guidance," "Structural Planning," "Structural Activation," "Generalization and Evaluation," and "Perspective."

The guidance function is carried out by counselors and counselees working cooperatively to solve anxiety producing problems so that the counselees can, with increasing skill, solve their own problems in a way to give meaningful, rational, and purposive serenity in the light of their own values. "Effective guidance, then, seeks not only to change the behavior that is essential to the solution of the immediate problem, but also to reorient the entire problem-solving behavior of the client" (p. 177).

Principles for accomplishing this goal which are elaborated and illustrated in the later sections are: analyzing, structuring and abstracting, observing real and symbolic experiences, using measurement of psychological behavior, implementing insight with structural planning and structural activation, handling the problem of values, getting beyond the talk stage, making the generalization of problem solving effective, and evaluating (as a counselor) one's own ability to bring about behavior change.

Readers of *Educational and Psychological Measurement* may have more than usual interest in the measurement and evaluation portions of the book. There is no itemized list of useful tests nor any evaluation of specific psychological measuring instruments. Rather the contributions relate quite specifically to problems of validity, reliability, and interpretation. For example, it is pointed out that test scores are not the same as the client's response repertory, but only a sample; and how representative we seldom know. Both a client's own description of his experiences and diagnostic test results "... are symbolic structures representing but not duplicating objective reality." The objectivity of the so-called *objective tests* and even *objectively derived test scores* is rejected. "These concepts are figments of intentional fantasy, for the selection of the items and the determination of the appropriate responses depend in a large measure on the attitudes of the test maker, while interpretation of the test score is so deeply imbedded in the symbolic reality of the counselor that its contact with objective reality may be a tenuous one" (p. 83).

It is pointed out that although psychologists and workers in related fields have been constructing psychological measuring instruments for nearly three quarters of a century there are only a few which are capable of supplying reasonably useful information about

behavior. Thus Weitz writes: "... the majority of psychometric instruments appear to give the user the impression that he is observing behavior when he is seeing only the fluttering of his own eyelashes in the viewpiece" (p. 98).

Why is there such a paucity of measures of real value? Because, this author holds, those who have been engaged in constructing such instruments have more often than not focused their attention on the response repertory, especially the communicative repertory, rather than on the behavior product. The entire individual's environmental context, of which the test is an important part, has usually been ignored.

Although these appear to be severe criticisms of present psychometric devices, the judgment is made that tests give more information about certain psychological behaviors than can be secured by other means. The need to improve these measures rather than to discard them is advocated.

One suggestion for such improvement is to investigate more thoroughly possible uses of situational observations, task situational tests, psychodrama, and devices for observing physiological mechanisms. It is held that these approaches to psychological measurement will force us into the use of a language closer to objective reality than such vocabulary as percentiles, T-scores, standard error of estimates, coefficients of validity and reliability, and norms. Also it is imperative to try to invent psychological measurement procedures which do not contribute so heavily to the very behavior which they are used to measure as do many present procedures.

At times the author seems somewhat hypercritical of educational situations and conditions. For example, "The melancholy facts of the matter appear to be that most of what we know as education, most contrived learning, takes place outside the school" (p. 47). Why melancholy, even if true? Or again, "We are frequently informed that what is required to force guidance into a more intimate alliance with perfection is more research—more and better. This melancholy adoration of research confronts us at every turn of a professional journal's page. It takes the form of ritualistic hymns extolling the virtues of research as the solitary road to salvation" (p. viii).

Perhaps these occasional emotionally laded statements are introduced for the purpose of achieving one of the author's objectives, namely, to "serve as goads to the reader, both practitioner and scholar, to stimulate him to some of this free-wheeling speculation, which the field of guidance so sorely needs" (p. x).

On the whole, the book develops a closely-knit set of principles around a somewhat new set of psychological concepts which is stimulating. The book may be productive of clearer understanding, better counseling, improved measurement and evaluation, and even

creative research, which could be based on this theoretical formulation for change through guidance.

CHESTER O. MATHEWS

University of California, Santa Barbara

Development in Early Childhood by D. Bruce Gardner. New York: Harper and Row, Publishers, 1964. Pp. x + 358.

The author has written an unusual book in the field of early childhood in the sense that the approach consists in a novel combination of themes. His first section, "How We Study Children," begins with a brief history of studies and attitudinal orientations toward children under the headings of (1) the behavioristic approach, (2) the normative-descriptive approach, (3) field theory, and (4) the psychoanalytic approach.

He defines child development in the context of a new field of integrated knowledge which is dependent upon several other areas of study for its strength, and cites specific contributions made to it by disciplines such as biology, psychology, sociology, and anthropology. In his chapter on the observation of children he suggests that information about children may be obtained on two levels: one, which is based on scientifically controlled settings, and two, informal types of observations which depend for pertinent analysis on the background and personal insight of the observer. This chapter could have been expanded. What was written was well done; in this particular field, however, increased attention is needed in the direction of sensitive and sophisticated interpretation of children's behavior.

Part Two, "Foundations of Development," is mainly a discussion of the orderly progression of growth of children from birth to six years old. Changes in size, strength, abilities, and language development are viewed not only in the biological sense but also in their relationships to the social environment. Beginnings of self-awareness are touched upon through discussions of Erikson's "sense of trust" learned by infants, and Harlow's experiment with monkeys which demonstrated the importance of gratification. Studies by Ribble, by Spitz, and by Bowlby stress the necessity for consistency in warm mother-infant relationships; a recent study by Gardner, Hawkes, and Burchinal indicated that infants are sturdier than has been commonly thought and that their fundamental needs may be cared for by a variety of mother figures rather than only one.

Chapter 7, "Communicating with Others," reports some studies on language development of children and on the ways in which the use of language is encouraged or hindered by the child's home environment. The relationship of language to the development of intelligence is brought into the following chapter which is a discussion of intellectual growth. The author defines intelligence as "...

the efficient operation of a complex of functions involved in a wide range of problem solving activities. Some aspects of intelligence are primarily verbal while others are mostly nonverbal. The ultimate criterion of intelligence is a social one: the ability of the child to adapt successfully to his world. In this concept of intelligence, however, there is room for creativity and inventiveness, as well as mere adaptation to the status quo" (p. 221). On the basis of recent research and current interest in creative abilities, definitions of intelligence have become broader than those given in earlier studies.

The emotional development of children is treated in Chapter 9. Gardner refers to three main processes in relation to emotional behavior: "(1) the physiological changes within the body, including changes in circulation, respiration, glandular activity, and sensory processes; (2) changes in the observable behavior pattern of an individual, such as laughing, crying, fighting, frowning, being 'silly,' or being moody; and (3) changes in the conscious experience of emotional awareness, which represent the 'feeling' component of emotion. The distinction among these three can be kept in mind by noting that the first can be measured with sensitive apparatus and electronic equipment, and the second can be observed by anyone. But the third, is available only to the individual experiencing it, and can be described only indirectly to another person" (p. 225). Studies dealing with emotional reactions of children are discussed, and the author presents in enjoyable style the ways that children learn to deal with various emotional experiences and conflicts.

Chapter 10 which discusses the process of achieving selfhood is the heart of the book for the author. He had stated in his preface that one of his objectives in writing the book was "to provide a fairly concise picture of the young child as he goes about his important task: *growing* . . . But there is a more fundamental emphasis which has less to do with age, *per se*: the big job of growing is a *job of achieving selfhood*. This is more than merely the theme of the book; it is the theme of the child. Every child's striving, seeking, playing, working, smiling, crying, hitting, running, and wondering are variations on that theme" (p. ix). His chapter centers around three major tasks in developing self-hood: ". . . learning to live in a world of tools, learning to live with other people, learning to live with oneself" (p. 270).

His concluding section, "The Society of the Preschool Child," is primarily an emphasis on the socialization process. Subcultural differences are described and explained in the light of the effects they have on the values of parents as they interact with their children. Family patterns emerge not in a vacuum but in culturally prescribed fashion; socio-economic background influences parental child-rearing methods in terms of authority and control, and the preferred qualities parents want to develop in their offspring. In Chapter 14,

"The Child and the Society," the author presents several perspectives of societal mores and their influences in the facilitation or inhibition in the pursuit of self-realization. "Another view (the one here held) [Gardner's parentheses] is that the function of suppressing and curbing the natural tendencies of the child is, at best, only one aspect of the relationship between child and society. The major aspect of the relationship is society's responsibility for bringing the child to the fruition of his own resources. The child is not only the product of his genetic inheritance, he is also the product of his social world—the society which he inherits. But, *the society is also the product of its children*" (p. 342).

The importance of certain value preferences in relation to a self-image and to one's orientation to the world is linked to the child's perception of the behavior of family members. Gardner states that the manner in which people do things around a child is considered to the child the *only* way; the child is not aware of alternative value systems. Thus the child's world is classified and defined, letting him know how he "ought" to perceive the world.

The author's particular style of bringing together materials which examine with heightened sensitivity the differences in children's orientation to their world (or *Weltanschauung*) and the value patterns related to the family's subcultures reflects the kind of insight that is needed in the study of child development.

EDYTHE MARGOLIN

University of California, Los Angeles

Empathy by Helen Barshay. New York: Exposition Press, 1964. Pp. 305. \$6.00.

Dr. Barshay, a counseling psychologist in a veterans' neuropsychiatric hospital, is concerned about the constricting elements which arise with feelings of hostility and aggression—feelings which consequently may cause the personality to close itself off from life. The preoccupations, the biases, and the extreme prejudices which accompany unwillingness to examine another person's point of view block creativity in the individual.

The three main objectives of the book seem to be: (1) an examination of the American value system where certain types of target people and specific organizations are concerned, (2) a consideration of those attitudes, which, once examined, may help society to understand and to aid those who need support in the amelioration of their problems, and (3) a creation of empathy which may allow hate and hostility to subside. Fulfillment of this goal will maximize the creativity and self-fulfillment of those who try to understand.

The author presents several chapters on stereotypical attitudes held by Americans in the late 1950's and early 1960's and asks that those views be re-evaluated. She discusses attitudes toward the

blind, the epileptic, and other physically handicapped individuals, the social deviant, the gifted, and the retarded—all suffer to some degree because the reactions of most people toward them do not facilitate favorable self-concepts.

The ability to relate to others is heightened by "empathy—the sensitivity and imagination to feel as others feel" (p. 25). This kind of "psychic nearness" (p. 25) is necessary in order to help others; it simultaneously enlarges the self as well.

The agitations, worries, and uncertainties which flourish in a society of wars, atom bombs, and rockets do not produce individuals who are not wholly trusting of their environment. But a dynamic kind of democracy can, says Dr. Barshay, produce socially mature individuals who are flexible and outgoing rather than power driven, aggressive, and selfishly immature. The destructiveness of hate has serious implications not only for mental health, but also for productiveness and self-fulfillment. Love is as essential as food—"Empathy is in fact a morality of love" (p. 26).

A chapter on attitudes toward family problems leads into a discussion of the close relationship between a lack of parental self-acceptance and disparagement of the child. Wholesome child-rearing practices which may result from realistic self-acceptance of the parents can produce a dependable, self-directing, and responsible adult. The democratic home life with greater emphasis on worthiness of human beings and less on criticism shapes the loving, trusting, and creative individual.

The chapters on attitudes toward labor-management problems, toward religion, and toward national and racial tensions emphasize the importance of a rationale behind the examination of those problems. An effective understanding becomes the basis for dealing with the mediation of conflict. Empathy replaces defensiveness.

The author cites an anonymous rhyme which suggests the nebulousness, yet misdirecting biases for labeling or diagnosing illness.

Psychosis is defined in terms of neurosis,
And neurosis is defined in terms of psychosis.
But please keep in mind that both are defined
In terms of the "normal" which is still undefined. (p. 15)

In her request that attitudes be re-examined, Dr. Barshay is essentially asking that people move beyond the point of accepting general attitudes toward atypicality or conflict; she hopes that they will be willing to generate some understanding through personal self-reflection in relation to the problems of others. Stereotypical societal images are accepted by the uninformed; this acceptance not only perpetuates fallacious images but also stultifies mental growth and self-fulfillment.

Although the passages in the foreword of the book present em-

pathy as apparently a new idea in psychotherapy, this reviewer has been under the impression that psychotherapists have been working for sometime in an empathic context with their patients. In essence, the author's call for the breaking of stereotypes is highly sound, but she has been neither alone nor unprecedented in her concern for the relationship between hostility and misguided attitudes.

Contemporary emphasis on the word "creativity" (the modern day catalytic agent) has endowed once more a classic and enduring concept with another facet such that the ability to identify with others now leads, instead of to an ordinary approach to self-understanding, to a creative one as is implied in the book's subtitle, "A Creative Approach to Self-Understanding."

Despite the presence of erroneous conceptions about the originality inherent in the selection of a discussion on stereotypical thinking, the book reflects a dignified tone and a sincere concern for the topic.

EDYTHE MARGOLIN

University of California, Los Angeles



<i>Situational and Individual Determinants of Attitude Scale Responses.</i> MARION STEININGER	757
<i>The Development of Personality Factors in Children and Adolescents.</i> MICHAEL S. BLACK	767
<i>Child Behavior Ratings: Further Evidence of a Multiple-Factor Model of Child Personality.</i> JOHN M. DIGMAN	787
<i>Acquiescence in the MMPI?</i> LEONARD G. RORER AND LEWIS R. GOLDBERG	801
<i>Acceptance of Sc Scale Statements by Visual Art Students.</i> IRWIN J. GOLDMAN	819
<i>The Validity of the Edwards Personal Preference Schedule (EPPS) Employing Projective and Behavioral Criteria.</i> DANIEL V. CAPUTO, JON M. PLAPP, CONSTANCE HANF, AND ANNE SMITH ANZEL	829
<i>Vocational Preference Patterns of Communications Graduates.</i> ALLEN E. IVEY AND MARK B. PETERSON	849
ELECTRONIC COMPUTER PROGRAMS AND ACCOUNTING MACHINE PROCEDURES	857
BOOK REVIEWS	893

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Contributors receive one hundred reprints of their articles without charge. Manuscripts should be sent in duplicate to G. Frederic Kuder, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia.

Subscription rate, \$10.00 a year, domestic and foreign. Single copies, \$2.50. Back volumes: Volume V or later, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Editor: G. Frederic Kuder

Associate Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

WILLIAM V. CLEMANS
Science Research Associates, Inc.

LOUIS D. COHEN
University of Florida

HAROLD A. EDGERTON
Performance Research, Incorporated

MAX D. ENGELHART
Chicago City Junior Colleges

E. B. GREENE
Chrysler Corporation

J. P. GUILFORD
University of Southern California

JOHN A. HORNADAY
Houghton Mifflin Company

E. F. LINDQUIST
State University of Iowa

FREDERIC M. LORD
Educational Testing Service

ARDIE LUBIN
U. S. Naval Hospital, San Diego

SAMUEL MESSICK
Educational Testing Service

WILLIAM B. MICHAEL
*University of California,
Santa Barbara*

HOWARD G. MILLER
*North Carolina State University at
Raleigh*

P. J. RULON
Harvard University

C. L. SHARTLE
Ohio State University

KENDON SMITH
*The University of North Carolina
at Greensboro*

THELMA G. THURSTONE
*University of North Carolina at
Chapel Hill*

HERBERT A. TOOPS
Ohio State University

JOHN E. WILLIAMS
Wake Forest College

E. G. WILLIAMSON
University of Minnesota

DOROTHY ADKINS WOOD
*University of North Carolina at
Chapel Hill*

VOLUME TWENTY-FIVE, NUMBER THREE, AUTUMN, 1965



CORRELATION AS A FUNCTION OF PREDICTOR SCORE POINTS

NAMBURY S. RAJU AND ISAIAH GUTTMAN

Science Research Associates, Inc.

THE basic Pearson product-moment correlation formula may be expressed as $r_{xy} = \Sigma xy / N$, where x and y are distributed with a mean of 0 and σ of 1. In this form, it is apparent that the correlation is an average function of score cross-products associated with each person. However, in a personnel selection validity situation, such as establishment of cutting scores, there may be a greater requirement to express the correlation as a function associated with individual predictor score points.

Several indices of predictive validity at different score points have been suggested. When the criterion is continuous, Berkson (1947) and Brogden (1949) proposed variations of the biserial correlation between the criterion and groups above and below any cutting score. When the criterion is dichotomized (e.g., satisfactory versus unsatisfactory performance), Richardson (1944) and Guilford (1950) proposed constructing a four-fold table (i.e., predicted satisfactory or unsatisfactory versus actual satisfactory or unsatisfactory) at any score point and calculating a phi coefficient. Berkson (1947), Guttman (1960), and Higgins (1963) proposed evaluating predictive validity in terms of the relative proportions of correct and incorrect predictions, and Guttman (1963) showed that there is a direct relation between the size of the phi coefficient and the difference in proportions.

The purpose of this paper is to define the validity correlation between a quantitative predictor and a quantitative criterion in terms of phi coefficients at each predictor cutting score point.

Derivation

Given a set of predictor scale values c which range from 0 to a maximum of T . Let c_i = i th cutting score (i.e., all persons at and below that score are rejected) and range from 0 to the maximum score minus one ($T-1$). At each c_i , assign a score 0 or 1 to each person: 0 if c_i is above his predictor score, 1 if c_i is below his predictor score. The predictor score for each person then equals the sum of his 0 and 1 scores across all c_i . For example, person A with a predictor score of 25 will get scores of 1 for each c_i from 0 to 24 (inclusive) and scores of 0 for each c_i 25 and higher, thus accounting for a predictor score of 25.

Let y = criterion scores, dichotomized,

x_p = predictor score for person p ,

C_{ip} = score for person p at cutting score i (0 or 1 score)

σ_x = standard deviation of predictor scores,

σ_{c_i} = standard deviation of the predictor score at each $c_i = \sqrt{p'q'}$,

σ_y = standard deviation of the criterion scores = \sqrt{pq} .

Then,

$$X_p = (C_{0p} + C_{1p} + \dots + C_{(T-1)p}), \quad (1)$$

$$X_p = \sum_{i=0}^{T-1} (C_{ip}), \quad (2)$$

$$r_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}. \quad (3)$$

Using equation (1), we may write equation (3) as

$$r_{xy} = \frac{Cov(c_0 + c_1 + \dots + c_{(T-1)})y}{\sigma_x \sigma_y}, \quad (4)$$

$$r_{xy} = \frac{r_{c_0y}\sigma_{c_0}\sigma_y + r_{c_1y}\sigma_{c_1}\sigma_y + \dots + r_{c_{(T-1)}y}\sigma_{c_{(T-1)}}\sigma_y}{\sigma_x \sigma_y}. \quad (5)$$

Since y is the criterion, σ_y is a constant at each c_i . Factoring out σ_y from equation (5),

$$r_{xy} = \frac{r_{c_0y}\sigma_{c_0} + r_{c_1y}\sigma_{c_1} + \dots + r_{c_{(T-1)}y}\sigma_{c_{(T-1)}}}{\sigma_x}, \quad (6)$$

$$r_{xy} = \frac{\sum_{i=0}^{T-1} r_{c_iy}\sigma_{c_i}}{\sigma_x}. \quad (7)$$

The formula given in equation (7) expresses the point-biserial correlation between the predictor and the dichotomized criterion as a function of phi coefficients at each cutting score. The $r_{c_i y}$ are the phi coefficients between the predictor, dichotomized at the c_i points, and the dichotomized criterion. The σ_{c_i} are the standard deviations at the C_i points, and equal $\sqrt{p'q'}$ (the square root of the product of the proportions above and below the c_i).

Equation (5) also yields the relationship shown in equation (8), that the covariance between the predictor and criterion equals the sum of covariances at each c_i .

$$r_{xy}\sigma_x\sigma_y = \sum_{i=0}^{r-1} r_{c_i y}\sigma_{c_i}\sigma_y. \quad (8)$$

REFERENCES

- Berkson, J. "Cost-utility as a Measure of the Efficiency of a Test." *Journal of the American Statistical Association*, XLII (1947), 246-255.
- Brogden, H. E. "A New Coefficient: Application to Biserial Correlation and to Estimation of Selective Efficiency." *Psychometrika*, XIV (1949), 169-182.
- Guilford, J. P. *Fundamental Statistics in Psychology and Education* (2nd Edition). New York: McGraw-Hill, 1950.
- Guttman, I. "Application of a Mini-Max Principle to Cutting Score Determinations." In *Tri-Service Conference on Selection Research*. ONR Symposium Report ACR-60, 1960, (pp. 111-117).
- Guttman, I. "A Minimum Loss Function as Determiner of Optimal Cutting Scores." A paper read at the American Psychological Association Convention, Philadelphia, Pa., 1963.
- Higgins, C. "Multiple Predictor Score Cut-offs Versus Multiple Regression Cut-offs in Selection of Academically Talented Children in Grade 3." In *Twentieth Yearbook of the National Council on Measurement in Education*. East Lansing, Mich., 1963, (pp. 153-164).
- Richardson, M. W. "The Interpretation of a Test Validity Coefficient in Terms of Increased Efficiency of a Selected Group of Personnel." *Psychometrika*, IX (1944), 245-248.



PREDICTING ACHIEVEMENT IN NINTH-GRADE MATHEMATICS FROM MEASURES OF INTELLECTUAL- APTITUDE FACTORS¹

J. P. GUILFORD, RALPH HOEPFNER, AND HUGH PETERSEN

University of Southern California

DURING the past 15 years, the Aptitudes Research Project (ARP) at the University of Southern California has demonstrated a great variety of new intellectual-aptitude factors, such as Thurstone called "primary mental abilities" (Guilford and Hoepfner, 1963). A natural implication from these findings is the question whether or not these unique abilities have significant relationships to areas of intellectual functioning in everyday life. Although the factors and their organization in the structure-of-intellect model have lent themselves to the development of psychological theory (Guilford, 1961, 1962, 1964), claims to social usefulness require other kinds of support. Factor analysis may be said to have demonstrated construct validity for the factors, thus satisfying the theoretical use of the factor concepts; information in the form of predictive validity is needed to satisfy suggestions of social significance.

As a general strategy, in the educational setting, it would be desirable to determine which of the many intellectual abilities are involved in successful mastery of any school subject, not only that we may better predict achievement or lack of it but also that teach-

¹ Based in large part on a project supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education, and Welfare, previously reported in full (Petersen, et al., 1963), with additional work supported by the Office of Naval Research, Personnel and Training Branch, under Contract Nonr-228(20). This material may be reproduced for any purpose of the United States Government.

ers, and even students, may become aware of the intellectual functions involved. The subject of study in the investigation reported here is ninth-grade mathematics, with the recognition that this is a critical place in the curriculum and that the usual academic-aptitude tests have not been as successful in predicting achievement in this subject as in some others.

It would have been possible to proceed in this study by trying out a large number of new tests, paying no attention to underlying constructs in the nature of dimensions of aptitude. Such a starkly empirical approach, which is recommended by Ghiselli (1964), was rejected, for, at best, in that approach one proceeds with only superficial knowledge, and at worst, in that direction lies chaos. There are hundreds of tests, but a limited number of aptitude factors. More important, the aptitude factors have psychological meaning, making possible the understanding of a criterion of achievement in terms of stable concepts in a general frame of reference. Thus, one can keep himself oriented in logically meaningful ways.

The use of tests of structure-of-intellect (SI) abilities seemed a promising approach for two reasons. One is the general expectation, based upon multiple-regression principles, that adding tests of entirely new factors to existing predictors of achievement would offer the greatest promise of improvement. One does not expect to increase multiple correlations simply by adding more tests of the same factors already included in a predictive battery. The second reason is that the ARP had distinguished a large category of abilities for dealing with symbolic information, apart from verbal and non-verbal (figural) abilities already recognized. It is fairly obvious that algebra capitalizes upon symbolic information and prominent among the abilities for learning it and for operating it must surely be symbolic factors.

The Problem and Hypotheses

The high school in which subjects were found for this study² taught four different mathematics courses, recognizing different levels of aptitude for algebra among ninth-grade students. The major distinction was between General Mathematics and Algebra,

² We are very much indebted to the teachers and students of the Lynwood, California High School for their generous cooperation in this project.

with two courses in each category. Some of the lower-aptitude students took a course in Basic Mathematics, which dealt with advanced arithmetic, with the introduction of equations and just a few algebraic concepts. Others took a course on Non-college Algebra which went a little further into algebra but the course was not considered college preparatory. In the next higher group the course was called Regular Algebra. The highest group took Accelerated Algebra, which extended into intermediate algebra.

The validation problem thus involved these four different courses, within each of which predictions of achievement constituted a goal. There was naturally interest in whether the pattern of predictive factor measures would be different depending upon the particular course. One might expect achievement to depend upon somewhat different intellectual abilities, at least in arithmetic versus algebra. Since the school had an administrative problem of classifying students in the four courses, year to year, a secondary interest in this study was to see whether factor tests could be used to discriminate successful algebra students from successful general-mathematics students, using composites of factor tests as a basis.

Fortunately, for the sake of a more thorough investigation, the school had routinely administered to the students three standard test instruments: the California Test of Mental Maturity (CTMM), the Differential Aptitude Test (DAT), and the Iowa Test of Basic Skills (Iowa), either at the end of the eighth-grade year or at the beginning of the ninth-grade year. Selected parts of these tests that were considered relevant were used in the study, both in factor-analytic operations and in multiple-regression studies. We were thus able to determine whether the new factor measures contribute as much or more to prediction of achievement and whether the one source of prediction added significantly to prediction by the other when they were combined.

Development of Hypotheses

The major hypotheses in this study dealt with forecasts as to which of the SI abilities should have the greatest promise of predicting achievement in general mathematics and algebra. There were two main sources of these hypotheses. One was an examination of four previous investigations that had considered the prob-

lem of the connection between intellectual factors and mathematical achievement (Canisia, 1962; Kline, 1956; Weber, 1953; and Werdelin, 1958). The other source was the SI model, and a consideration of which abilities seem to be most represented in what the student actually has to do in comprehending algebraic and arithmetical ideas and in solving such problems.

The four investigators just mentioned factor analyzed quite a variety of tests, and as nearly as one can tell, in all four studies combined, about 26 of the SI abilities were involved. The tests had been selected because they were thought to have some relevance in connection with mathematics. Three of the authors made some attempts at predictive validations, but the net result seemed to be that only measures of the two factors of *verbal comprehension* and *general reasoning* seemed to have much to offer by way of prediction. These two factors probably dominate the verbal and quantitative types of scores commonly used in academic-aptitude batteries. From an examination of the probable factors involved in these four studies, it is easier to see why their new kinds of tests did not contribute to prediction. Of the 26 factors, only 8 were composed of symbolic (letter or number) content; the great majority were either verbal (semantic) or figural. Except for one of the studies, which used tests composed of algebra problems as criteria (Kline, 1956), criteria were not carefully chosen.

Hypotheses from the Structure of Intellect

It is perhaps necessary to remind the reader concerning the major concepts and categories involved in the SI model. Table 1 lists the three sets of factor categories and their letter symbols.

TABLE 1
*Sets of Categories for the Three Dimensions of the SI Model and
Their Letter Symbols Used in Combinations of Three to Identify Each Ability*

Operations	Contents	Products
Cognition.....C	Figural.....F	Units.....U
Memory.....M	Symbolic.....S	Classes.....C
Divergent production.....D	Semantic.....M	Relations.....R
Convergent production.....N		Systems.....S
Evaluation.....E		Transformations.....T
		Implications.....I

Each intellectual ability is classified in one category from each list, its descriptive label specifying its kind of *operation*, its kind of *content*, and its kind of *product*. Each ability has its own unique conjunction, symbolized by a combination of three letters. Thus, the trigram MSI stands for "memory for symbolic implications," a factor more commonly known as "numerical facility." The factor pertains to implications because it is the salient factor in tests devoted in large part to numerical operations. An operation like $2 \times 5 = 10$ includes the given information " 2×5 ," which should imply 10. It is as if the examinee who performs thus is saying "If 2×5 , then 10," to put the statement in its complete form. It is a matter of memory because such implications have been practiced and retained. It is a matter of symbolic information because the operations are concerned with numbers, not verbal meanings or concrete objects, which are semantic and figural contents, respectively.

It has already been stated that the abilities involving symbolic information were regarded as most promising in connection with mathematics. The SI model has places for 30 such abilities, with five kinds of operation combined with six kinds of products, as shown in Table 2. The trigram symbols are given for abilities that

TABLE 2

A Matrix of the Structure-of-Intellect Abilities Pertaining to Symbolic Information

		Operations				
Products	U	CSU	MSU	DSU		ESU
	C	CSC		DSC		ESC
	R	CSR*	MSR	DSR*	NSR*	ESR**
	S	CSS*		DSS**	NSS*	ESS**
	T				NST*	EST**
	I	CSI*	MSI*	DSI**	NSI*	ESI**

* A factor of primary interest.

** A factor of secondary interest (see text).

have been demonstrated at the time this report was written; six of these had not been demonstrated at the time the mathematics-validation study was initiated but were being investigated concurrently with the mathematics study. With the aid of a few examples, we shall now see how some of the factors are logically involved in certain aspects of algebra.

The cognitive abilities, in the first column of Table 2, are prob-

ably more involved in understanding information as given by the text or by the teacher, but they should also function in understanding the nature of given expressions, equations, and problems. CSR, the cognition of symbolic relations, should come into play in recognizing the relations in such expressions as a/x , a^3 , and $a = xy$. In each case, elementary symbols are related to one another in certain ways. CSI, the cognition of symbolic implications, involves recognition that one expression follows from another, for example, that $a^3 = a \times a \times a$, and that if $a + y = 7$, then $y = 7 - a$.

The divergent-production factors have to do with generating a number of ideas from given information; it is something more than understanding, for it involves retrieving known information from memory storage. In each case, alternative ideas, or multiple answers, are produced. DSS is the divergent production of symbolic systems. A typical kind of system in algebra is a complex expression or an equation that has some degree of complexity. Asking a student to write several different examples of a quadratic equation should be a task involving the factor DSS. Factor DSI, the divergent production of symbolic implications, should be involved in tasks calling for multiple inferences. For example, given the two simple equations:

$$D + 2 = C - 7$$

$$D + C = 2X$$

how many new simple equations can the student generate, such as $D = C - 9$, $D = 2X - C$, and $2(X - C + 1) = -7$?

Convergent production is like divergent production in that the answers must come from the individual's own stock of stored information, but instead of several alternative answers being acceptable there is only one possible right answer. Enough information is given to fully determine the answer. Factor NSS, the convergent production of systems, should apply where the student must perform operations in a fixed sequence in order to solve a problem. In one of the tests of NSS, the examinee is told to start with the number 5 and arrive at 1, using only the steps: $\times 2$, $\div 7$, and -3 ; his task is to state the order that will make the transition. Order is one kind of system; an organized sequence. The ability NST, convergent production of symbolic transformations, should apply where the student revises an expression, as in simplifying or factoring.

Evaluation means inspecting and judging whether information is "good," adequate, or sound. For factor ESS, evaluation of symbolic systems, we might give an item such as:

Which expression has the same form as $(4x^2 + 9)$?

The alternatives: (A) $(x^2a^2 + b)$ (B) $(3a^2 + 4)$ (C) $(9x^2 - 8)$

From Table 2 it can be seen that all abilities having to do with units and classes of information were regarded as less promising for attention in a predictive-validation study. This is not to say that any one of those abilities is regarded as totally irrelevant; the scope of the study had to be limited, because of limitations on testing time; only a limited number of factors could be brought knowingly into the study. It may be that the abilities dealing with classes will prove to be of importance in the "new" mathematics; the students of this study, in 1961, were studying traditional mathematics.

Table 2 indicates nine symbolic factors of primary interest, which means that each of them was represented in this study by two tests so that factor analyses could be done with data, to determine whether the tests were measuring the factors as expected. Three semantic factors (not shown in Table 2) were also included among those of primary interest, two of which were CMU (verbal comprehension) and CMS (general reasoning), since they dominate standard academic-aptitude tests. Instruction in mathematics is largely verbal, hence these two factors should be involved. Furthermore, when the student translates verbally stated problems into equation form, he must first understand the problem as a semantic system (hence CMS involvement) before he can produce the equation (hence NSS involvement). A third semantic factor included among those analyzed, although the factor had not been previously demonstrated, was EMR, evaluation of semantic relations, for which tests had been developed in another study. The battery of tests for the factors of primary interest required eight hours' testing time, being administered near the beginning of the school year.

Later in the school year, two additional hours of testing time became available. This time was devoted to tests of the factors designated as being of secondary interest in Table 2. The tests had been developed for some concurrent studies, which have now been completed and which demonstrated that the abilities for which

these tests were designed can be differentiated (Gershon, et al., 1963; Hoepfner, et al., 1964). These tests were not analyzed in connection with the mathematics study, since there was only one test designed for each factor, except for factor ESS. Attention to four of the evaluation abilities indicates some belief that the student must check his own work, as he goes along and as he rejects and accepts results of operations he has performed. The attention to divergent-production abilities was in recognition of the fact that much trial-and-error behavior occurs in solving mathematical exercises. From the SI point of view, trial and error is a matter of divergent production alternated with evaluation.

The Test Variables

Since the test variables are the empirical referents for the factors and represent the factors in the prediction studies, it is necessary to give the reader at least the minimum information regarding them. In what follows, the tests are listed in alphabetical order (within sets), for easy reference, each with a code number and a line of description. The first three letters of the code number in each case indicates the factor for which the test was intended and in most cases the test had been previously found loaded on the factor, usually in adult populations. The tests did not always prove to be measures of those factors in the two analyses made with the ninth-grade subjects in this study. Following each test variable derived from standard aptitude tests, the salient factors they are hypothesized to measure are indicated in parentheses. For more complete information regarding the tests, including distribution and reliability statistics, see Petersen, et al. (1963) or Guilford and Hoepfner (1963), both of which indicate the sources of the tests. The ones not analyzed in this study were subsequently analyzed by Nihira, et al. (1964) and Hoepfner, et al. (1964).

1. Alternate Additions—DSRO1B. Show how numbers in a set may be related in obtaining the same total, in different ways.
2. Best Trend Name—EMRO1A. Choose the word that best describes the order of four given words, the order indicating a variable, such as time.
3. Camouflaged Words—NSTO1A. Find within a meaningful sentence a group of consecutive letters that, in the given order, spells the name of a sport or a game.

4. Circle Reasoning—CSSO1C. Discover the common principle by which one circle is blacked in each of four rows of mixed circles and dashes.
5. Correlate Completion II—NSRO1A. Supply a word that bears the same relation to the single word as the relation between the words in two given pairs, the relation being based on letters rather than meaning.
6. Form Reasoning—NSI02B. From a table of equations involving geometric forms as symbols, solve some other equations involving the same forms.
7. Letter-Number—NSIO3A. Find the relationship between letters and digits and use the relationship to find the number that corresponds to a new letter.
8. Letter Series—NSRO2B. Find the rule of order in a series of letters, then fill in a blank with the letters that would fit the rule.
9. Letter Triangle—CSSO2A. With letters arranged systematically within a triangular pattern, which letter should appear in a marked, vacant place?
10. Matched Verbal Relations—EMRO2A. Choose one of four pairs of words that has the same relation between words as that of the given word pair.
11. Necessary Facts—CMSO4A. Determine what information is needed to attain solutions for given arithmetic problems in which needed facts are missing.
12. Number Rules—DSRO2B. Starting with a given number, arrive in several different ways at a second given number, applying a single arithmetical operation.
13. Numerical Operations, Guilford-Zimmerman Aptitude Survey, Part III—MSIO1A. Simple operations with numbers.
14. Picture Arrangement—NMSO2A. Given four pictures from a comic strip in scrambled order, put them in the correct temporal sequence.
15. Right Order Test—NSSO1B. Starting with one given number, do three given numerical operations in the right order to obtain a second specified number.
16. Seeing Trends II—CSRO1A. Describe a trend in a series of words, where a certain letter relation determines the trend.
17. Sentence Order—NMSO3A. Arrange three given sentences in a sensible temporal order.

18. Ship Destination Test, Form A-2—CMS. State how many miles a ship travels from one point to another, considering such variables as distance, directions, wind, current, and starting position.
19. Sign Changes—NSIO1A. Solve simple arithmetic equations in which the operation signs are to be changed according to rules.
20. Symbol Grouping—CSIO1A. Rearrange scrambled symbols in a specified systematic order as efficiently as possible.
21. Word Changes—NSSO2B. Given a set of words, one designated as first and one as last, arrange the remaining words in proper sequence so that only one letter is changed in going from one to the next.
22. Word Linkage—EMRO3A. Choose from a list of three words the one that is related to two given words by virtue of two different meanings.
23. Word Patterns—CSIO2B. Arrange a list of short words efficiently in a kind of crossword-puzzle design.
24. Word Relations—CSRO2B. Recognize the same relation between words in each of two pairs, then complete a third pair using the same relation.
25. Word Transformations—NSTO2A. Regroup the letters of words in a phrase so as to make another phrase.
26. CTMM—Language MA (CMU).
27. CTMM—Non-Language MA (MSI and CMU).
28. Iowa Reading Comprehension—TEST R. (CMU).
29. Iowa Arithmetic Concepts—Test A-1. (CMU and MSI).
30. Iowa Arithmetic Problem Solving—Test A-2. (MSI and CMS).
31. DAT—Verbal Reasoning. (CMU, CMR, and NMR).
32. DAT—Numerical Ability. (MSI).
33. DAT—Abstract Reasoning. (CFR).
34. DAT—Clerical Speed and Accuracy. (ESU)
35. Sex membership.
38. Abbreviations—ESIO1A. Choose the word that the given abbreviated word most likely implies.
39. Condensations—ESTO1A. Choose the better of two shorthand alternatives, choice to be based on the unique meaning of the shorthand.
40. Letter-Number Scales—ESSO1A. Given the numerical values

of two letters of the alphabet, estimate the numerical value of a third letter, from three alternatives, none of which may be exactly correct.

41. Limited Words—DSIO2B. Given two words, make up additional pairs of words (anagrams) using all the letters in the given pair and no others.

42. Most Similar Sets—ESSO2A. Choose which of two sets of letter-number sets is most like the given set.

43. Number Combinations—DSSO2B. Write several different equations using only the given numbers and given operation signs.

44. Sign Changes II—ESRO1B. Make the sign changes necessary to make an expression into an equation.

There were two criterion tests, constructed especially for this study, one in general mathematics and one in algebra, constituted so as to sample systematically achievement in the two kinds of courses according to the stated objectives for those courses. Course grades were also used as separate criteria, but were found to be less consistently predictable than the achievement-examination scores. Only the special-examination scores will be considered as criteria for this report.

Procedure

Test Administration

The battery of 25 factor tests that were to be factor analyzed was administered October 10, 11, 12, 1961, to the entire ninth-grade class, which numbered approximately 600, by ARP personnel with assistance from the school. The secondary battery of seven tests was administered in the mathematics classes May 2-3, 1962. Testing conditions appeared to be good. The achievement tests used for criterion measures were administered May 23-24 in regular class sessions, after nearly a full school year of class instruction.

Factor Analyses

Four factor analyses were carried out. Two were in samples of boys and girls separately, to determine whether sex differences would affect the factor structure. In these analyses, varimax rotations were made and all possible pairs of factors after rotation were compared by means of the Tucker criterion of factor con-

gruence (Tucker, 1951). There was evidence of fairly satisfactory similarity of the two factor structures, so two other analyses were carried out, in the general-mathematics groups and the algebra groups, respectively, with sexes combined.

For these two analyses, three score variables were added from the standard aptitude tests in order to help determine factors CMU and MSI, making 28 factored variables. The N 's for the two samples were 217 and 211. Fourteen principle-axes factors were extracted. Varimax rotations failed to yield satisfactory solutions for purposes of interpretation, so further orthogonal rotations were made graphically.

One reason for the factor analyses was to be able to obtain a set of factor scores, in order to achieve a much smaller number of predictor variables for the multiple-regression analysis. To obtain factor scores, the simple procedure used was to sum the standard scores of the tests for each factor, where more than one test had a substantial or higher loading on the factor. No test was used in more than one factor composite. The tests entering into factor scores differed in some places for the general-mathematics versus the algebra students, but every factor had at least one test in common to the composites for the two groups. One factor was residualized in rotation, leaving the basis for 13 factor scores. Such scores should have more intercorrelation than would the factors, but as it turned out, the highest inter-factor-score correlation was .50.

Multiple Correlations

Multiple correlations and other multiple-regression statistics were computed with each of the achievement-examination variables as dependent variables. Combinations of independent variables included: the 13 factor scores, the 7 additional test scores, and a combination of the two sets (20 predictor variables); 2 CTMM scores, 3 selected Iowa test scores, 4 selected DAT scores, and a combination of the 9 standard-aptitude-test scores. In order to determine whether the factor scores would contribute significantly to prediction over and above that available from the standard aptitude tests, the 13 factor scores were combined as predictors along with the CTMM scores, the Iowa scores, and the DAT scores, with F tests to determine significance of gains in multiple correlation.

A second set of multiple-regression treatments applied a step-wise correlation program, which starts with the best predictor then adds the next best in turn, making an F test to determine whether the increase in R is significant at each step. In the results given in this report, predictors from the 9 standard tests and the 20 factor variables were treated separately. In either case, the list of predictors in each of the four courses was cut off when F went below that needed for significance at the .10 level.

For none of the multiple-regression analyses were crossvalidation studies made. With the regression equations limited to each course, the numbers of cases were limited to 73 in the smallest group to 101 in the largest.

Discriminant Analysis

To determine whether combinations of factor tests or scores can discriminate well between successful algebra students and successful general mathematics students, procedures were used described by Tiedeman, et al. (1951) and by Cooley and Lohnes (1962). "Successful" was defined as being above the median in the achievement examination for the group. This analysis was not done in terms of the 13 factor scores for the reason that by using standard scores in combining test scores, as specified by the computer program that was available, means were automatically equated for the two groups, so raw scores from the 25 factor tests were used in one analysis and from the 7 other tests in a second analysis. The 9 most discriminating tests from both sources were then used in a third analysis, which, of course, gave biased results.

Results

The statistical results concerning means, standard deviation, reliability estimates, intercorrelations, and factor matrices will not be presented here. They are available in a much more comprehensive report (Petersen, et al., 1963). The data reported here pertain to selected multiple-regression results and discriminant-function analysis.

Table 3 presents the multiple correlations between weighted combinations of parts of the three standard academic-aptitude tests (separately and in total combination), the 7 factor tests, the 13 factor scores, and the combination of the 20 factor measures, in

TABLE 3

*Multiple Prediction of Mathematical-Achievement Scores from
Weighted Composites of Standard Tests and of Factor Tests**

Prediction Battery	Basic Mathematics	Non-college Algebra	Regular Algebra	Accelerated Algebra
9 standard tests	.60	.53	.22	.74
2 CTMM tests	.34	.40	.18	.37
3 Iowa tests	.53	.31	.20	.62
4 DAT tests	.57	.53	.24	.70
7 factor tests	.42	.56	.27	.51
13 factor scores	.46	.45	.39	.75
20 factor predictors	.48	.54	.38	.74

* The coefficients of multiple correlation given are unbiased, i.e., corrected for shrinkage.

each of the four courses. All multiple R 's have been corrected for bias, to take into account the numbers of predictors.

From the first and last rows of Table 3 it appears that the two kinds of composites, the 9 standard predictors versus the 20 factor predictors, just about break even. The standard predictors apparently do better in predicting achievement of basic mathematics students; the factor battery apparently does better in the regular algebra group. In the latter group, predictions were consistently poorer than elsewhere, possibly indicating something wrong with the measurement of achievement in that group. A more realistic comparison might well be made between results for each standard instrument taken separately and the factor batteries, since most schools presumably depend upon a single aptitude instrument. In this kind of comparison, we see that the combination of the four DAT tests does about as well as the larger factor batteries, but the CTMM and the Iowa composites apparently do less well. The factor battery of 20 variables undoubtedly takes more time to administer than the four DAT tests, also the battery of 13 tests, but if the non-predictive factor tests were eliminated it might well be found that the factor battery would be more efficient in terms of testing time, for each factor test is relatively short.

With respect to the regression equations for the factor variables, we can gain some impressions as to the relative importance of different factors and categories of factors. Most of the following impressions arise from examination of the beta weights which were reported in full by Petersen, et al. (1963). First, it can be said that there is a slight tendency for more factors to be relevant in connection with Accelerated Algebra than for other courses. Second,

the factors dominant in most standard academic aptitude tests, CMU and CMS, are conspicuous for their absence in most regression equations. The exceptional success of the four DAT scores is apparently attributable to their emphasis on other, non-verbal, factors. A fair guess would be that the DAT Numerical emphasizes factor MSI, which is strongly represented in all equations except that for Accelerated Algebra. The DAT Abstract emphasizes factor CFR, which has been found in another analysis done by the ARP (O'Sullivan, et al., 1965). DAT Clerical emphasizes the factor ESU, for which there is indirect evidence (Hoepfner, et al., 1964). From the beta weights for the DAT tests, the MSI and ESU scores had the most consistently higher contributions to make in the four courses. The role of factor ESU in this study is a bit curious, since Osborn and Melton (1963) did not find the DAT Clerical score to be predictive.

A third general impression is the apparent relative importance of evaluation factors in some of the equations. The role of ESU was demonstrated by the performance of the DAT Clerical test. Sign Changes II, which was previously found to be loaded on a factor identified as ESR, was more recently found heavily loaded on factor ESC, evaluation of symbolic classes (Hoepfner, et al., 1964). If classes are, indeed, the products most heavily involved in this test, it would suggest that success in algebra depends in part upon making fine distinctions regarding classes or types of expressions and equations. The tests for factors ESS and EST, which show betas that merit attention in some equations, have not yet been analyzed, but that they belong in the evaluation category leaves little doubt, by virtue of their similarity in form to other evaluation tests that have been analyzed.

Another noteworthy impression is the relative lack of cognitive tests with substantial betas. By actual count, there were 40 opportunities for cognitive variables to show substantial betas but they did so in only four instances. Possibly the right cognitive factors were not represented, but it was surprising that the factors CSR, CSS, and CSI, which *were* represented, did not show up better in the predictions. The absence of CMU and CMS from the list of relevant factors has already been commented upon. The cognitive factors might be expected to play greater roles during the initial stages of learning, in the students' following mathematical instruc-

tion; weaknesses in these respects might be compensated for in various ways during a course. But there would seem to be plenty of room left for individual differences in understanding problems given in items of an achievement examination.

Of special interest are the apparent roles of the few divergent-production factors represented in the study—DSR, DSS, and DSI. Analyses have demonstrated the reality of these dimensions of intellect. At least one of these factors had some indications of predictive contribution in all four courses. The student's facility in producing a variety of alternative answers or approaches to problems evidently contributes to his success in ninth-grade mathematics of all kinds. Other divergent-production factors not studied in this connection, such as the unknown factor DST, might also add something of value.

Prediction from Combinations of Standard and Factor Variables

The fact that the list of factor variables includes many factors not represented in the standard aptitude tests, and the fact that both kinds of composites are highly predictive of achievement, suggest that adding the factor variables to the standard-test variables would yield even higher levels of prediction. Table 4 presents

TABLE 4
Increases in Multiple Correlation from Adding 13 Factor Scores to Each of Three Standard Batteries of Academic-Aptitude Tests, and F Ratios for Testing Significance of Increases

Prediction Battery	Basic Mathematics		Non-college Algebra		Regular Algebra		Accelerated Algebra	
	R	F	R	F	R	F	R	F
CTMM (2 scores)	.35		.41		.21		.38	
CTMM + 13 scores	.59	1.58	.59	1.60	.54	2.25*	.80	6.06**
Iowa (3 scores)	.55		.34		.24		.63	
Iowa + 13 scores	.65	1.05	.58	1.94*	.54	2.10*	.82	3.23**
DAT (4 scores)	.59		.55		.29		.72	
DAT + 13 scores	.64	0.46	.59	0.48	.55	2.07*	.85	3.36**

* Significant at the .05 level; ** significant at the .01 level.

a summary of the evidence that this is so. Additions of the 13 factor variables was made to each of the three standard batteries used. Although the multiple *R*'s are noticeably increased as a result of the addition of the factor variables in all cases, the increases were

not significant in Basic Mathematics, and mostly not significant (2 cases out of 3) in Non-college Algebra. In general, the higher the level of the course, the larger and more significant are the gains. The gains are least for additions to the DAT battery and most for additions to the CTMM battery, as should be expected from the results in Table 3. The addition of the seven factor variables might have yielded further increases in the R 's, since other factors were involved in them. At any rate, the principle of improving prediction by bringing new factors into the regression equations is well supported.

Reduced or Stepwise Regression Equations

We next consider the stepwise multiple-regression results, in which only those predictors were retained, in either the standard test batteries or the factor batteries, that made statistically significant contributions (.10 level) to prediction. The results are given in Table 5, for the four courses separately and for combinations of the two general-mathematics and the two algebra groups, since in either case the students had the same achievement examination. Here the nine scores from the standard batteries were placed in competition with one another for inclusion of their parts.

Among the standard tests, DAT Numerical led the list, by far, in most groups, consistent with its probably heavy saturation with factor MSI, which was strongly relevant also in most of the factor-variable equations. The DAT Abstract variable appears in three of the lists, indicating that factor CFR is relevant. The appearance of Iowa Reading Comprehension in results for the Accelerated Algebra group and in results for the combined general-mathematics group is surprising, in view of the fact that it was expected to be largely a CMU measure. But reference to its factor loadings explains its appearance among the leading variables, in view, also, of the absence of a more purely CMU test (Verbal Comprehension) from the lists. In the algebra group, analysis showed that the Iowa Reading Comprehension test had just about as high a loading in factor EMR as in CMU. Reference to the factor list for Accelerated Algebra shows EMR next to the top. In the general-mathematics group, Iowa Reading Comprehension had as high a loading for NMS as for CMU. Both Reading Comprehension and factor NMS appear in the lists of significant predictors in the Basic

TABLE 5

Tests Contributing Significantly (F at the .10 level) to Multiple Predictions, and Multiple Correlation Coefficients, for Weighted Composites of Standard Aptitude Tests and of Factor Tests and Some of Their Factor Composites

Course	Standard Aptitude Test Scores		Factor Composites and Test Scores		
	Test	R	Variable	R	N
Basic Mathematics	DAT numerical	.59	MSI	.59	77
	Iowa Read. Comp.		DSI		
Non-college Algebra	DAT numerical	.49	NMS		
	DAT abstract		ESR	.62	95
			NSR		
			(ESS)*		
			MSI		
Regular Algebra	DAT numerical	.29	EMR		
	CTMM Non-lang.		DSR	.45	101
			MSI		
Accelerated Algebra	Iowa Read. Comp.	.76	NSR	.78	73
	DAT numerical		EMR		
	DAT clerical		DSR		
	DAT abstract		NSS		
General Math.			(ESS)		
	DAT numerical	.65	MSI	.67	173
	Iowa Read. Comp.		ESR		
	DAT abstract		NSR		
Algebra	DAT clerical		DSS		
	DAT numerical	.58	DSR	.65	174
	Iowa Read. Comp.		MSI		
	CTMM Non-lang.		NSS		
	DAT clerical		EMR		
			NSI		
			ESI		

* Factor symbols in parentheses indicate that the test was designed for the factor but it has not been analyzed.

Mathematics group but neither appears in the lists for the Non-college Algebra group. The variable of CTMM Non-language appears in the algebra groups, but nothing can be said about its probable factor contribution, since this variable was not factor analyzed.

The relevant factors in the lists from the factor variables is much the same as previously discussed. Symbolic factors dominate the scene, so far as content is concerned, and evaluation factors are in the majority so far as operation categories are concerned, with some divergent- and convergent-production factors represented and the prominent memory factor MSI usually present.

The multiple correlations are fairly close to those obtained from

the corresponding complete lists of variables. In general, as before, the R 's are generally similar for combined standard versus factor batteries, with the factor batteries having superiorities in two of the courses and in the combined algebra group. The combined algebra group showed smaller R 's than the Accelerated Algebra group alone but higher R 's than the Regular Algebra group alone, as should be expected.

The Discriminant Analyses

Three discriminant analyses were made, one involving the 25 factor tests that entered into the factor analyses, one the seven other factor tests, and one with a combination of the nine most discriminating contributors from the two sources. It would have been better to use the 13 factor scores in the first instance, but, as indicated earlier, those scores had been scaled within the two groups (general mathematics and algebra, respectively). Consequently raw scores were used, with some of the results being given in Table 6. Only tests that contributed .03 or more toward discrimination are listed there. The F ratio was significant at the 0.1 level. Making actual classifications in terms of the discriminant function with 25 predictors, there were only 12 false negatives and 12 false positives. The phi correlation between predicted and actual group memberships was .77. For the seven factor tests three discriminated with contributions greater than .03, with only nine misclassified students in each group and a phi of .83.

The selection of the best discriminators for the third analysis gives biased results, of course, and the use of an F ratio is very questionable, since the selection of variables was not random. With only seven misplaced students in each group, the phi coefficient is still .83. The results, in general, show what kind of success one may expect in discriminating two such groups of students using factor tests. Discriminations available from the standard tests were not studied, with the expectation that, being more complex factorially and thus intercorrelating more, they would do much less well. The classification of students in the two groups had also been based in part on information from one or more of those standard instruments.

It is of some interest to consider what factors are most relevant for this kind of prediction. If a factor measure predicts achieve-

TABLE 6

Discriminant-Function Solutions Comparing 103 Successful (Above-Median Achievement) Algebra Students with 105 Successful General-Mathematics Students, Using Weighted Combinations of 25, 7, and 9 (Selected) Tests, with Proportions of Contributions to Composite Discrimination

Test Number and Name	Factor	Proportion of Contribution	F Ratio	Phi Coefficient
With 25 tests*			8.99*	.77
3. Camouflaged Words	NST	.031	(12 students	
7. Letter-Number	NSR	.030	misclassified in	
11. Necessary Facts	(CMS, CMU) ^b	.031	each group)	
12. Number Rules	DSR	.040		
15. Right Order Test	NSS	.055		
20. Symbol Grouping	CMS	.032		
With 7 other tests*			47.07*	.83
41. Limited Words	DSI	.246	(9 students	
42. Most Similar Sets	(ESS) ^c	.207	misclassified in	
43. Number Combinations	DSS	.090	each group)	
With 9 selected tests			48.80*	.83
3. Camouflaged Words	NST	.054	(7 students	
7. Letter-Number	NSR	.046	misclassified in	
11. Necessary Facts	(CMS, CMU) ^b	.049	each group)	
12. Number Rules	DSR	.037		
15. Right Order Test	NSS	.036		
20. Symbol Grouping	CMS	.040		
41. Limited Words	DSI	.235		
42. Most Similar Sets	(ESS) ^c	.228		
43. Number Combinations	DSS	.061		

* Only tests having a proportion of contribution greater than .03 are given.

^b Salient factors differ in the two groups of students, that for algebra students being given first.

^c A test designed for factor ESS but not yet analyzed.

* F is significant beyond the .01 level.

ment within each of the two kinds of groups to be discriminated (total general mathematics and total algebra groups), we should not expect it to be necessarily discriminating between the two groups. Discrimination depends upon difference between means, however, so there might still be room for both kinds of prediction from the same factor variable. But, in general, we should expect a somewhat different list of significant predictors in the lists for making the two kinds of predictions, between versus within groups.

This expectation is partially born out. For example, factors CMU and CMS do make some small contributions to discriminations where they fail to predict much within groups. This is probably because those factors were weighted in making the original classification of students, using standard tests. Factor NST, convergent production of symbolic transformations, appears in the

discrimination lists but not in the achievement-prediction lists. Four other factors (DSR, DSS, ESI, and EMR) are apparently relevant for discrimination but for prediction in either general mathematics or algebra, only, not in both. Several factors appear as predictors in Table 5 but do not appear as discriminators in Table 6, including MSI, NMS, EMR, AND ESI. All these differences suggest that exactly the same test battery could not be best used for doing the double duty of classifying students and also predicting achievement in the two kinds of course. There could be much in common, however, for factors DSR, DSS, DSI, NSR, NSS, NSI, and ESR are represented in both lists. It is noteworthy that production abilities dominate this list. It is also noteworthy that two tests stand out well above the others in proportion of contribution to discrimination, namely, Limited Words, for factor DSI, and Most Similar Sets, designed for factor ESS but unanalyzed as yet. Within the context of the factors involved in this study, it appears that 10 would need to be considered for a predictive battery, and perhaps 3 additional tests to cover the task of classification, although the latter, for factors CMU, CMS, and NST, might not add enough to pay for their use.

Discussion

It is hoped that the report of this validation study demonstrates the value of keeping informed concerning construct validity while investigating predictive validity, a principle and a strategy that the senior author urged a number of years ago (Guilford, 1948). The reasons will not be repeated here, except to say that this approach satisfies one's scientific curiosity, to the extent that factor information is available, as one proceeds, and there are possibilities of drawing more general conclusions of theoretical and practical importance. The only deterrent to such an approach should be the extra work involved. With high-speed computers available, such resistance should become minimal.

It should also have been demonstrated how multiple predictions of achievement in certain courses of study can be greatly increased by bringing together within the predictive battery a number of factors, many of which have never been utilized before for the same purpose. A pattern has been set for the investigation for any other course of instruction. The major limitations still remaining are needs for comprehensive and carefully considered criterion

measures and the lack of measures for many undemonstrated factors that are hypothesized by the SI model. The latter deficiency will be remedied in time; of the 120 intellectual abilities indicated by the present model, it can be said that 70 have been demonstrated and 15 undemonstrated factors are currently under investigation. The criterion measures will have to be tailored to fit the particular course situation.

In general, it is urged that course grades be passed by as potential criteria, unless it can be shown that they have real promise of validity in the light of course objectives, and consistency of value within sets of the sample of students in the validation study. The correlations between grades and the achievement-test scores in the study reported here were in the neighborhood of .50. Had there been higher correlation, grades would have been given more attention in this investigation.

The question naturally arises as to whether predictions of achievement in the "new" mathematics courses will take the same kinds of predictor variables as have been found in the traditional types of courses. Osburn and Melton (1963) found that predictions in the two kinds of courses, new and traditional, were about the same, using such tests as the Iowa Algebra Aptitude Test, the Orleans Algebra Prognosis Test, the Thurstone PMA tests, and the DAT battery, but there was some evidence of interaction of ability and type of course. With a much larger variety of abilities represented, it is likely that more evidence of differential validities would be found.

Summary and Conclusions

The purposes of this investigation were (1) to see whether composites of measures of structure-of-intellect factors would yield a high degree of prediction of achievement in courses in ninth-grade mathematics, compared to three test batteries of the more traditional, standard, type of academic aptitude tests; (2) to see whether such measures would contribute predictive value over and above that available from standard test batteries; (3) to see how well a battery of factor tests can discriminate between successful students in general mathematics and in algebra; and (4) to see which factors are most relevant for predictions of success within courses and between courses.

After factor analyses of 28 test variables, 13 factor scores were

derived from combinations of tests, and 7 additional factor tests were employed, involving other factors. Predictor scores were available from three commonly used academic-aptitude batteries, nine such scores in all. Four mathematics courses ranging from Basic Mathematics (advanced arithmetic) to Accelerated Algebra (intermediate algebra) with approximately 400 students who completed all tests were available. The criterion measures were especially designed final examinations of two kinds—general mathematics and algebra.

The major conclusions are as follows:

1. Batteries of factor scores were better predictors of achievement than two of the standard-test combinations, especially in the prediction of achievement in algebra.

2. A composite of 13 factor scores gave increased prediction when added to each of the three standard-test combinations, significantly so in the algebra courses.

3. Combinations of factor-test scores discriminated between successful (above-median) algebra students and general mathematics students with an accuracy close to 90 percent.

4. With only predictors that gave statistically significant contributions to prediction of achievement, some 12 different factors were found relevant. Most of these factors are from the symbolic category of the structure of intellect; very few are cognition factors and quite a number are evaluation factors; most of them deal with the products of relations and implications.

5. Of the factors that are relevant to discrimination between the two kinds of students (general mathematics and algebra), most of them are also relevant for prediction within courses of either kind or both, thus much the same kind of battery could be used to make both kinds of discrimination.

REFERENCES

- Canisia, Sister M. "Mathematical Ability as Related to Reasoning and Use of Symbols." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 105-127.
- Cooley, W. W. and Lohnes, P. R. *Multivariate Procedures for the Behavioral Sciences*. New York: Wiley, 1962.
- Gershon, A., Guilford, J. P., and Merrifield, P. R. "Figural and Symbolic Divergent-Production Abilities in Adolescent and Adult Populations." *Reports from the Psychological Laboratory*, No. 29. Los Angeles: University of Southern California, 1963.
- Ghiselli, E. *Theory of Psychological Measurements*. New York: McGraw-Hill, 1964.

- Guilford, J. P. "Factor Analysis in a Test-Development Program." *Psychological Review*, LV (1948), 79-94.
- Guilford, J. P. "Factorial Angles to Psychology." *Psychological Review*, LXVIII (1961), 1-20
- Guilford, J. P. "An Informational View of Mind." *Journal of Psychological Researches*, VI (1962), 1-10.
- Guilford, J. P. "Intelligence, Creativity, and Learning." In R. W. Russell (Ed.), *Frontiers in Psychology*. Chicago: Scott Foresman, 1964, pp. 125-147.
- Guilford, J. P. and Hoepfner, R. "Current Summary of Structure-of-Intellect Factors and Suggested Tests." *Reports from the Psychological Laboratory*, No. 30. Los Angeles: University of Southern California, 1963.
- Hoepfner, R., Guilford, J. P., and Merrifield, P. R. "A Factor Analysis of the Symbolic-Evaluation Abilities." *Reports from the Psychological Laboratory*, No. 33. Los Angeles: University of Southern California, 1964.
- Kline, W. E. "A Synthesis of Two Factor Analyses of Intermediate Algebra." *Technical Report*. Princeton, N. J.: Educational Testing Service, 1956.
- Nihira, K., Guilford, J. P., Hoepfner, R., and Merrifield, P. R. "A Factor Analysis of the Semantic-Evaluation Abilities." *Reports from the Psychological Laboratory*, No. 32. Los Angeles: University of Southern California, 1964.
- Osburn, H. G. and Melton, R. S. "Prediction of Proficiency in a Modern and Traditional Course in Beginning Algebra." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIII (1963), 277-287.
- O'Sullivan, Maureen, Guilford, J. P., and de Mille, R. "Measurement of Social Intelligence." *Reports from the Psychological Laboratory*, No. 34. Los Angeles: University of Southern California, 1965.
- Petersen, H., Guilford, J. P., Hoepfner, R., and Merrifield, P. R. "Determination of 'Structure-of-Intellect' Abilities Involved in Ninth-grade Algebra and General Mathematics." *Reports from the Psychological Laboratory*, No. 31. Los Angeles: University of Southern California, 1963.
- Tiedeman, E. V., Rulon, P. J., and Bryan, K. G. "The Multiple Discriminant Function—a Symposium." *Harvard Educational Review*, XXI (1951), 71-95.
- Tucker, L. R. "A Method for Synthesis of Factor Analysis Studies." *Personnel Research Selection Report*, No. 984. Washington, D. C.: Department of the Army, 1951.
- Weber, H. "Untersuchungen über de Faktorenstruktur numerischer Aufgaben." *Zeit. f. esp. u. angew. Psychol.*, 1953, 3.
- Werdelin, I. "The Mathematical Ability." *Investigationes IX, Studia Psychologia et Paedagogica*. Lund, Sweden: 1958.

DIFFICULTY AND OTHER CORRELATES OF CRITICALNESS RESPONSE STYLE AT THE ITEM LEVEL¹

LAWRENCE J. STRICKER

Educational Testing Service

LARGELY as a result of Cronbach's (1946; 1950) reviews of the early literature on response sets—or response styles—on tests, the view is widely held that tests and test items that are difficult or ambiguous are most affected by those response styles, such as acquiescence, evasiveness, and extremeness, that are linked to the response format. Despite the prevalence of this view, the results of relevant studies are not altogether convincing, and are largely limited to two response styles—acquiescence, and the position of the chosen alternative on multiple-choice tests.

One major group of relevant studies examined acquiescence on tests or items of varying difficulty or ambiguity. Only one of the studies (Gage, Leavitt, and Stone, 1957) concerns the difficulty issue. In a comparison of difficult and easy general information items, the reliability of the number of *true* responses was .68 for the 50 difficult items, and .09 for the 40 easy ones. The two kinds of items, however, did not represent exactly the same content areas.

The other studies of this kind concern various facets of item ambiguity. Two of these studies directly concern ambiguity. In

¹ This study was supported by the Office of Naval Research, under Contract Nonr-2338(00). The author wishes to thank Dr. Norman Frederiksen for suggesting this study, and Miss Henrietta Gallagher for supervising the computations.

Tables reporting the means and standard deviations of the item variables and all their intercorrelations for each test have been deposited with the American Documentation Institute. Order Document No. 8453, remitting \$1.25 for 35-mm microfilm or \$1.25 for 6 by 8 in. photocopies.

one study (Bass, 1955), the original and reversed version of the California *F* scale items that were rated most opposite in meaning (and, hence, assumed to be clear in meaning) appeared to measure content (i.e., the correlation between the two items was relatively high and in the content direction), while those rated less opposite in meaning (and, hence, assumed to be relatively ambiguous in meaning) did not, evidently because of the operation of acquiescence on the latter items. Moreover, an analysis of variance found significantly ($p < .01$) greater acquiescence to the pairs of items that were least opposite in meaning. The assumption in this study that the rated level of opposition measures ambiguity is questionable, for it is possible for a pair of items that are rated as similar in meaning to be very unambiguous. In a second study (Banta, 1961), the number of *agree* responses on attitude scales increased and the number of extreme responses decreased with the ambiguity of the scales' referents. The referents were "President Eisenhower," "College Fraternities," and "People in General"—the first referent was presumed to be the least ambiguous and the last referent the most ambiguous. Unfortunately, the content of the items on the scales also varied with the referent, so the effects of these two variables are confounded.

Two other studies investigated the readability of items, which should be one cause of ambiguity. In one study (Stricker, 1963), more *agree* responses and fewer socially desirable responses were made to hard- than to easy-to-read attitude items, but fewer *agree* and fewer socially desirable responses were made to hard- than to easy-to-read personality items. In another study (Hanley, 1962), the reliability of the number of *true* responses was generally higher for long MMPI items than short ones.

The other major group of relevant studies concern two other response styles on achievement or aptitude tests and examine the relationship between the subjects' performance on such a test and their tendency to make stylistic responses on the same test. These studies are analogous in their approach to the previous studies that compare tests or items of varying difficulty, for a test, by definition, is more difficult for those who perform poorly on it than for those who perform well.²

² It should be noted that three studies (Frederiksen, 1958; Frederiksen and Messick, 1959; Helmstadter, 1957), which employed one or more of three re-

On a pitch discrimination test that required the subject to respond *high* or *low* (Cronbach, 1946), one response style score based on the 20 most difficult items—the difference in the number of correct responses to items keyed *high* and items keyed *low*—had a reliability of .49 for the entire sample of subjects and .56 for those who made five or more errors on the 20 items. The significance of the difference between these two reliability coefficients was not reported.

Three studies of multiple-choice tests are relevant; all are concerned with a response style tendency linked to the location of the chosen alternatives. Two studies (Cronbach, 1950) concern the reliability of such a response style score—a tendency to choose alternatives located before rather than after the keyed alternative. In one study, which employed a modified version of the Ohio State University Psychological Examination, the reliability of the response style score was .20 for the entire sample of 171 subjects, .29 for the 65 lowest scoring subjects, and .54 for the 26 subjects with the lowest scores among the group of 65. The significance of the differences between these coefficients was not reported. In a second study, which employed the Henmon-Nelson Test of Mental Ability, the reliability of the response style score was .09 for 66 subjects who were wrong on 30 or more of the 90 items and .42 for another 84 subjects who were wrong on 50 or more of the items. These two coefficients are significantly different ($t = 2.1$, $p < .05$), as computed by the present author. A third study (Marcus, 1963) employed four multiple-choice tests, each being administered to a different sample of subjects. The tests contained the same 100 achievement items, but the correct alternatives and distractors appeared in a different position. Within each test, each of the four alternatives was correct an equal number of times. The number of responses made to the four alternatives was reported to differ sig-

port writing tests (Frederiksen, 1958), together with special scoring procedures (Helmstadter, 1957) that yield separate content and response style measures from them, have reported low negative correlations between these two measures. These studies were not intended to bear on the issue of the relationship between difficulty and response style and their results are not relevant, for the response style measure purposely reflects the direction of the response style as well as its strength—a high score indicates a stylistic tendency towards style as well as its strength—a high score indicates a stylistic tendency towards criticalness, a low score indicates a stylistic tendency towards criticalness, a score midway between these two indicates the absence of a stylistic tendency in either direction.

nificantly ($p < .05$) in the entire set of 100 items (though the most popular alternative varied with the test), but not in the 20 most difficult items. In addition, the second and fourth alternatives, though none of the others, were reported to differ significantly ($p < .05$) in the percentage of correct responses made to them in the entire set of items, but none of the corresponding differences in the difficult items were reported to be significant. It is difficult to appraise these results and the related significance tests, for they depend on nonindependent observations; the basic statistic in this study—the total number of responses—pools the number of subjects and the number of responses by each subject.

The present study was undertaken to clarify the link between difficulty and a response style of criticalness at the item level by examining the relationship between the difficulty of an item and the item's correlations with a score for this response style as well as a content score, both scores being based on all the items in the test that includes the item. In addition, for exploratory purposes, the relationships of these three item variables to item readability measures (Chall, 1958) and several characteristics of the item format were also appraised.

Method

Tests

Three report writing tests (Frederiksen, 1958)—Alternative Expressions Test, Recognizing Ambiguities Test, and Evaluation of Revisions Test—were used in this study. Each of the tests requires the subject to evaluate written passages and, hence, resembles ordinary aptitude or achievement tests. The characteristics of these three tests are as follows.

Alternative Expressions. There are 70 items in this report-writing test, each consisting of a sentence in which a word or phrase is underlined, followed by another word or phrase in parentheses. The task is to judge whether or not the word in parentheses could safely be substituted for the underlined expression. The instructions require the respondent to be rigorous, as though a policy decision, administrative action, or even a legal claim might hinge upon the interpretation of the sentence. The sentence is marked *Same* or *Dif-*

ferent, depending on whether such consequences are judged to be same or different. A sample item (keyed *Different*) is:

He implied that he would resign. (It could be inferred.)

Recognizing Ambiguities. The subject is required to judge whether each of the 50 items in this test is ambiguous or unambiguous. The instructions define "ambiguity" as allowing more than one grammatically defensible interpretation. Sample item (keyed *Ambiguous*):

The plot is pure farce, involving a phony British nobleman's quest for the hand of an American heiress, a social climbing American mother, and a visiting English lady named Mrs. Wollope.

Evaluation of Revisions. Each of the 40 items in this test consists of a short statement paired with a revision of the statement, such as might be prepared by an editor. The task is to decide whether the revision could safely be substituted for the original, i.e., whether the consequences would be the same no matter which version was used. The instructions state that the interpretation should be rigorous, as though an important decision or even a legal claim might hinge upon the choice of a passage. Sample item (keyed *Different*):

Original

John Worthy is recently reported to be under consideration for appointment to an undersecretaryship in the State Department. He is past president of the DRA.

Revision

John Worthy is a past president of the DRA and is under consideration for appointment to an undersecretaryship in the State Department, according to a recent report.

(Frederiksen and Messick, 1959, pp. 141-142.)

Separate content and response style scores for each of these tests were obtained with the Helmstadter (1957) scoring procedure, as modified by Messick (1961). The content scores can vary from -1 to $+1$, and the response style scores can vary from -1 to $+1$. In this study, a response style score of -1 represented a tendency to make "critical" responses (i.e., *Ambiguous* or *Different*), and a score of $+1$ represented a tendency to make "uncritical" responses (i.e., *Unambiguous* or *Same*).

Subjects

Two of these three tests were administered to each of three samples of students: (a) male and female undergraduates at Emory

University; (b) male students in a graduate school of business and (c) first-year students at the law schools of eight major universities.³

The Undergraduate and Business School samples each took the Recognizing Ambiguities Test and the Alternative Expressions Test, and the Law School sample took the Recognizing Ambiguities Test and the Evaluation of Revisions Test.

Subjects who did not reach the last item on both tests were excluded from the analysis. As a result, there were usable data 208 students in the Undergraduate sample, 320 in the Business School sample, and 400 in the Law School sample.

Procedures

The two tests in each sample were analyzed separately. Four basic statistics were computed for each item in each sample.

(a) The conventional difficulty index—the percentage of “correct” responses chosen by the entire sample of subjects.⁴ The correct response to each item was determined originally by “expert” keying.

(b) The modified difficulty index—the percentage of correct responses chosen by those subjects who predominantly responded on the basis of content. This index is intended to minimize the distortion that could be produced in the conventional difficulty index by a tendency to make stylistic responses, particularly if this tendency is independent of the items’ difficulty. Although this index is based on those subjects who respond on the basis of content, it is also the best available estimate of item difficulty for all subjects, including those who respond on the basis of response style. The tendency for the latter group to make stylistic responses interferes with a more direct measurement of the item difficulty for this group.

These indexes for each test were based on the subgroup of subjects with response style scores on the test of .10 to $-.10$, from

³ Thanks are due Dr. Frederiksen, Dr. John R. Hills, and Dr. Sam C. Webb for furnishing the completed answer sheets for these samples.

⁴ A transformed version of this difficulty index, obtained by converting the percentage choosing the correct response to a normal curve deviate, was also computed but is not reported because it correlated .99 or 1.00 with the original difficulty index in each of the tests in all of the samples.

rich enough subjects were randomly eliminated so that there were an equal number of subjects in the .01 to .10 and $-.01$ to $-.10$ intervals. These subgroups for the Recognizing Ambiguities Test consisted of 59 students in the Undergraduate sample, 94 in the Business School sample, and 117 in the Law School sample; the subgroups for the Alternative Expressions Test consisted of 56 students in the Undergraduate sample and 63 in the Business School sample; and the subgroup for the Evaluation of Revisions Test consisted of 77 students in the Law School sample.⁵

(c) The content validity index—the biserial correlation of the item with the content score. This index is a measure of the extent to which the item reflects content. The signs of these correlations were retained in subsequent analyses, but there were only seven instances in which an item had a negative correlation with the content score, the largest of these correlations was $-.05$, and in no instance did the same item have a negative correlation in more than one sample.

(d) The response style validity index—the biserial correlation of the item with the criticalness response style score. This index is a measure of the extent to which the item reflects this response style, in the sense that the chosen response alternatives are related to the overall tendency to make "critical" or "uncritical" responses. Only one of these correlations indicated that those who chose the uncritical alternative were more "critical," as determined by their overall response style scores, than those who chose the critical alternative.

In addition, several other item variables were computed or obtained. Four were available for all three tests: item location (its number), position of correct alternative (0-first; 1-second), mean word length (in syllables), and mean sentence length (in words). Total passage length (in words) was available for the Recognizing Ambiguities Test and the Evaluation of Revisions Test. The total length (in syllables) of the two key phrases and the total frequency of usage in general writing (Thorndike and Lorge, 1944) of the words in these phrases were available for the Alternative Expressions Test.

⁵ The transformed version of this difficulty index correlated .97 to .99 with the original version.

*Results**Independence and Reliability of Content and Response Style Scores*

In appraising the meaningfulness of analyzing the two sets of item validity indexes on each test, the independence of each test's two scores, from which the corresponding item indexes were derived, and the reliability of these scores were examined. Table 1 reports the correlation between the content and response style scores for each test in each sample. Two kinds of correlations are reported. One involves the scores based on all the items. The second is an estimate that is intended to eliminate any possible distortion in the first correlation produced by the experimental dependence of the two scores. This estimate was obtained by computing the mean, using a z transformation, of (a) the correlation between the content score for one half of the items (X items) and the response style score for the other half of the items (Y items) and (b) the correlation between the response style score for the X items and the content score for the Y items. The corresponding correlation for a full-length test was then estimated from this mean correlation, using the Spearman-Brown formula.⁶

Three of the six correlations between the content and response style scores based on all the items were significant ($p < .05$), but the largest correlation, disregarding sign, was only .13. None of the estimated correlations that were derived from the independent subsets of items were significant ($p > .05$).⁷

The corresponding split-half reliability coefficients for the content and response style scores appear in Table 2. The results for the various samples were similar. The reliability of the response style scores was moderate. It was somewhat higher for the Alternative Expressions Test (.74 and .76) than for the two shorter tests (their reliability coefficients range from .49 to .70). The re-

⁶ The X and Y subsets of items were so chosen that they have exactly the same proportion of items for which the first alternative is keyed correct and are roughly the same in their location in the test. These two subsets of items were also the basis of the split-half reliability coefficients reported in this study.

⁷ When these correlations were corrected for attenuation in both variables, five of the correlations based on all the items were significant ($p < .01$)—the largest of these correlations, disregarding sign, was .32—and one of the correlations based on the independent subsets of items was significant ($r = .18$, $p < .01$).

TABLE 1
Correlations between Content and Response Style Scores

Sample	N	Recognizing Ambiguities Test		Alternative Expressions Test		Evaluation of Revisions Test	
		Actual	Estimated	Actual	Estimated	Actual	Estimated
Undergraduate	208	-.03	-.05	-.11	.09	—	—
Business School	320	.06	.02	-.11*	.02	—	—
Law School	400	.12*	.03	—	—	.13**	.01

* Significant at .05 level; ** significant at .01 level.

TABLE 2
Split-Half Reliability of Content and Response Style Scores

Sample	N	Recognizing Ambiguities Test		Alternative Expressions Test		Evaluation of Revisions Test	
		Content	Response Style	Content	Response Style	Content	Response Style
Undergraduate	208	.37	.58	.33	.74	—	—
Business School	320	.26	.59	.23	.76	—	—
Law School	400	.41	.70	—	—	.34	.49

liability of the content scores was consistently and appreciably lower than the response style scores and roughly similar for the three tests. These reliability coefficients ranged from .23 to .41.

Stability of Item Indexes

It was possible to obtain lower-bound estimates of the stability of the item difficulty and validity indexes of the Recognizing Ambiguities Test and the Alternative Expressions Test because both tests had been administered in at least two of the samples, which differed markedly in their characteristics. The correlations between the same item indexes in the different samples appear in Table 3. The difficulty indexes were the most stable. The conventional difficulty indexes for the items of the Alternative Expressions Test correlated .69 ($p < .01$) in the two samples that took that test, and these indexes for the items of the Recognizing Ambiguities Test correlated .91 to .94 ($p < .01$) in the three samples that took it. The corresponding correlations for the modified difficulty indexes were .63 ($p < .01$) for the Alternative Expressions Test and .86 and .92 ($p < .01$) for the Recognizing Ambiguities Test. The validity indexes were appreciably less stable. The content validity indexes correlated .33 ($p < .01$) and the response style validity indexes correlated .34 ($p < .01$) in the samples that took the Alternative Expressions Test. The content validity indexes correlated .44 to .53 ($p < .01$) and the response style validity indexes correlated .16 ($p > .05$) to .40 ($p < .01$) in the samples that took the Recognizing Ambiguities Test.

Correlations between Item Indexes

The correlations of the items' difficulty indexes with their two validity indexes, which were computed separately for each test in each sample, appear in Table 4. Five of the six correlations of each of the difficulty indexes with the content validity indexes were significant, ranging from .32 ($p < .05$) to .57 ($p < .01$) for the conventional difficulty indexes, and .24 ($p < .05$) to .56 ($p < .01$) for the modified difficulty indexes. Only the correlations (.05 for the conventional difficulty indexes and .06 for the modified ones) for the Recognizing Ambiguities Test in the Business School sample were not significant ($p > .05$). None of the six correlations between the conventional difficulty indexes and the response style validity

TABLE 3
Stability of Item Indexes over Samples

Samples Compared	Recognizing Ambiguities Test (<i>N</i> = 50)				Alternative Expressions Test (<i>N</i> = 70)			
	Conventional Difficulty	Modified Difficulty	Content Validity	Response Style Validity	Conventional Difficulty	Modified Difficulty	Content Validity	Response Style Validity
Undergraduate and Business School	.91**	.86**	.44**	.40**	.69**	.63**	.33**	.34**
Undergraduate and Law School	.92**	.89**	.53**	.16	—	—	—	—
Business School and Law School	.94**	.92**	.51**	.34*	—	—	—	—

* Significant at .05 level; **significant at .01 level.

TABLE 4
Correlations of the Items' Difficulty Indexes with Their Validity Indexes

Sample	Recognizing Ambiguities Test (N = 50)		Alternative Expressions Test (N = 70)		Evaluation of Revisions Test (N = 40)	
	Content Validity	Response Style Validity	Content Validity	Response Style Validity	Content Validity	Response Style Validity
Undergraduate Business School Law School	Correlations with Conventional Difficulty Indexes					
	.32*	-.11	.38**	.06	—	—
	.05	-.03	.34**	-.17	—	—
Undergraduate Business School Law School	Correlations with Modified Difficulty Indexes					
	.34*	-.11	.34**	.09	—	—
	.06	.02	.24*	-.07	—	—
	.33*	-.10	—	—	.56**	-.35*

* Significant at .05 level; **significant at .01 level.

indexes were significant ($p > .05$), but one of the corresponding correlations for the modified difficulty indexes was significant—a correlation of $-.35$ ($p < .05$) for the Evaluation of Revisions Test in the Law School sample. The scatterplots for all these correlations were inspected; none of the regressions appeared to depart from linearity.

Moreover, very similar correlations were obtained for the two tests—Recognizing Ambiguities Test and Alternative Expressions Test—that were administered to two or more samples when one sample's difficulty indexes were compared with a different sample's validity indexes. (These correlations appear in the tables described in footnote 1.)

The correlations between the two validity indexes, which were computed separately for each test in each sample, and which appear in Table 5, were less consistent. Three of the six correlations, however, were significant, and each of these was negative. Moreover, each of the samples and each of the tests was represented in one of these three correlations— $-.40$ ($p < .01$) for the Recognizing Ambiguities Test in the Undergraduate sample, $-.31$ ($p < .01$) for the Alternative Expressions Test in the Business School sample, and $-.35$ ($p < .05$) for the Evaluation of Revisions Test in the Law School sample.

Table 6 reports the correlations of the item difficulty and validity indexes with the item location and position of correct alternative. None of the correlations of the item location with the two item difficulty indexes or the content validity indexes were significant ($p > .05$), but four of the six correlations of item location with the response style validity indexes were significant and positive. These significant correlations, which ranged from $.28$ ($p < .05$) to $.56$ ($p < .01$), occurred at least once in each sample and at least once for each test. The scatterplots for all the correlations with item location were inspected; none of the regressions appeared to depart from linearity.

None of the correlations between the position of the correct alternative and either of the two item difficulty indexes were significant ($p > .05$). The position of the correct alternative did have several significant correlations with the two validity indexes, all but one of which were positive in sign, indicating that the correlations with the total score were higher for those items in which the

TABLE 5
Correlations of the Items' Content Validity Indexes with Their Response Style Validity Indexes

Sample	Recognizing Ambiguities Test (<i>N</i> = 50)	Alternative Expressions Test (<i>N</i> = 70)	Evaluation of Revisions Test (<i>N</i> = 40)
Undergraduate	-.40**	-.03	—
Business School	-.05	-.31**	—
Law School	-.03	—	-.35*

* Significant at .05 level; **significant at .01 level.

TABLE 6
Correlations of the Items' Location and Position of Correct Alternative with Their Difficulty and Validity Indices

Sample	Recognizing Ambiguities Test (<i>N</i> = 50)			Alternative Expressions Test (<i>N</i> = 70)			Evaluation of Revisions Test (<i>N</i> = 40)		
	Conventional Difficulty	Modified Difficulty	Content Validity	Conventional Difficulty	Modified Difficulty	Content Validity	Conventional Difficulty	Modified Difficulty	Content Validity
Undergraduate Business School Law School	.03	.10	.34*	.16	.18	.01	.12	.08	.07
	.01	.03	.56**	.14	.12	.09	.28*	.05	.31*
	.03	.03	.25	—	—	—	—	—	—
Undergraduate Business School Law School	.06	.00	.02	.06	.01	.28*	.05	.01	.07
	.15	.01	.13	.21	.06	.27*	.27*	.01	.07
	.02	.02	.43**	—	—	—	—	.22	.57**

* Significant at .05 level; **significant at .01 level.

second alternative was keyed correct. Its significant correlations with the content validity indexes were .29 ($p < .05$) for the Recognizing Ambiguities Test in the Law School sample, .28 ($p < .05$) for the Alternative Expressions Test in the Undergraduate sample, and .27 ($p < .05$) for the same test in the Business School sample. Its significant correlations with the response style validity indexes were .43 ($p < .01$) for the Recognizing Ambiguities Test in the Law School sample, .27 ($p < .05$) for the Alternative Expressions Test in the Business School sample, and $-.57$ ($p < .01$) for the Evaluation of Revisions Test in the Law School sample.

None of the correlations of any of the five item readability indexes with the difficulty or validity indexes were significant ($p > .05$) for any test in any sample.

Discussion

The most important finding is that an item's difficulty is not related to the extent that the item measures the response style of criticalness; the choice of response alternatives is equally dependent for easy and difficult items on the subjects' overall response tendency to criticalness or uncriticalness. The inconsistency between this finding and the generalization that difficult items elicit response styles (Cronbach, 1946; Cronbach, 1950) may arise from differences between criticalness and the response styles investigated in the studies that were previously viewed as supporting this generalization, as well as the limitations of some of these studies.

This finding about item difficulty, as well as the finding that the item's readability is not related to the extent that the item measures criticalness response style, suggests the desirability of re-examining conceptualizations of response styles as phenomena that appear only in rather unstructured situations (Cronbach, 1946; Cronbach, 1950; Rapaport and Berg, 1955), or that are "emitted in the absence of any known stimulus" (Bass, 1957, p. 83). (Such a conceptualization underlies the derivation of the procedures used in obtaining separate content and response style scores in the present study.) At least some of the response styles may be so potent and pervasive that they exert their influence almost without regard for the properties of the test or test items. This view, which is consistent with the conceptualization of response styles as stable, individual-difference variables that reflect important aspects of per-

sonality (Jackson and Messick, 1958), suggests that research aimed at clarifying the meaning of response styles may be more fruitful if it examines the dynamics of the person rather than the characteristics of the item.

The results that concern the extent to which the item measures content are also interesting. The tendency for an item's difficulty to be related to the extent that the item measures content—the easy items being better measures—was consistent over tests and over samples, and also agrees with a previous finding about achievement tests (Brogden, 1956).

There are hints in the present data that the extent to which an item measures content and the extent to which it measures criticalness response style are negatively related, suggesting that a test item tends to measure one of these tendencies or the other, but not both. This possibility points up the potential feasibility of constructing relatively "pure" measures of each of these response tendencies by identifying those items that predominantly measure content and those that predominantly measure response style. These findings, as well as the finding that the validity indexes have some stability, also suggest that both validity indexes are predictable, even though only the content validity indexes were related to the difficulty indexes. Hence, it may be possible to identify other item variables that are associated with one or both of the validity indexes.

One such item variable is the location of the item in the test. The item's location was generally related to the extent that the item measures criticalness response style—the later the item in the test, the more it measures response style—although it was unrelated to the extent to which the item measures content. This relationship between item location and response style seems to be a causal one: the items on each test were not arranged in any systematic order, and item location was not related to any of the other item variables—item difficulty, the readability measures, and the position of the correct alternative (except for a significant correlation of .29 ($p < .05$) with the latter for the Alternative Expressions Test).

The link between item location and criticalness response style uncovered in this study represents a special instance of "secular trends" (Loevinger, 1957) on psychological tests—changes in score over time—which have been reviewed by Fiske and Rice

(1955) and Windle (1954). These reviews described a number of studies that found changes between two or more administrations of the same test when no experimental treatment intervened between the two administrations, and even when one administration immediately followed the other. The findings in the present study, as well as in earlier ones (Gordon, 1952; Mollenkopf, 1950; Whitcomb and Travers, 1957), are even more striking because they demonstrate changes within the same test, even a very short one. The present findings also agree with previous results that indicate that response styles may be associated with secular trends. Evidence already exists about social desirability response style. Windle's (1954) review indicated that more socially desirable scores on the standard content scales were generally obtained when personality inventories were administered a second time, and Gordon (1952) found that more socially desirable responses were made to the later items in a personality inventory. The role of the other response styles is less clear. Scores on an acquiescence response style measure, the Couch and Keniston (1960) Agreement Response Scale, did not change significantly ($p > .05$) over a three-month interval (Stone and James, 1961), and scores for extremeness response style did not change significantly ($p > .05$, computed by the present author) on the Perceptual Reaction Test or the Word Reaction Test, either for 7-day or 15-day intervals (Berg, 1953).

The reason for the present finding, much less the general phenomenon of secular trends, is not clear. One possible explanation is that these effects are caused by changes in the subjects' knowledge about the test (Eisenberg and Wesman, 1941; Kaufman, 1950; Whitcomb and Travers, 1957; Windle, 1955)—the purpose of the test, the nature of favorable performance on it, and means of attaining such performance. On the report writing tests, for example, the subjects, as a result of responding to the items, may begin to see and make finer, but no more valid, distinctions between the alternative words and passages. Similarly, on personality inventories, subjects, as they respond to the items, may realize that the items vary systematically in social desirability and proceed to respond in terms of this variable. A somewhat different explanation of these trends is that they are caused by changes in the subjects' approach to the test (Caldwell, 1959; Fiske, 1957; Fiske, 1961; Lighthall, et al., 1960; McCreary and Bendig, 1954; McGeoch and

Whitely, 1927; Mollenkopf, 1950; Sarason, et al., 1960; Voas, 1956; Windle, 1955)—the energy, interest, and anxiety elicited by the test. For example, subjects' energy or interest may decrease as they respond to the test items, and they may then tend to make responses consistent with their stylistic predilections, which requires less expenditure of energy than pondering the items before making responses to their content (Jackson, 1959; Stricker, 1963). Such a decrease in energy and interest might produce a less critical approach to the items of the report writing tests.

A more systematic investigation of the psychological meaning of individual differences in these trends may be rewarding. Such research on changes between two administrations of the same test has already begun (Caldwell, 1959; Windle, 1955), but a similar investigation of trends within tests is still lacking. Both kinds of studies might profitably examine the links of these trends to such variables as learning ability, defensiveness, energy level, and test-taking motivation.

A related question of some interest is whether the observed response trend within the test occurs for all subjects, regardless of the initial strength of their predisposition to make stylistic responses, or whether it is limited to those who are initially predisposed to make stylistic responses. The present results could have been obtained, even though the trend was not present in most of the sample, if it occurred in sufficient strength in the remaining fraction of the sample.

While the dynamics that underlie the relationship between item location and criticalness response style observed in the present study are not clear yet, some of its more important implications are apparent. This relationship lends support to Loevinger's (1957) criticisms of traditional psychometric theory for ignoring such secular trends in performance. At the same time, this finding, when considered together with the parallel finding of a lack of relationship between item location and content, suggests that many of the changes in test scores observed by Loevinger and others may occur because of changes in response style tendencies rather than content. Such a possibility can only be adequately appraised after research on this issue with such major response styles as acquiescence, social desirability, and extremeness.

The response style finding also has a number of implications for

the measurement of this variable and, perhaps, other response styles as well. Most generally, the intriguing possibility is suggested that, insofar as the items are affected by their sheer location, such item characteristics as reliability and validity are unknowable. The measurement of these properties requires that the item be in a particular location in the test, but this location affects these very same properties. In any event, the assumption that items near the beginning of a test and those near its end are equivalent is likely to be incorrect, especially for tests that are long and that measure such response styles. This assumption is explicit in such common psychometric formulas as the Spearman-Brown prophecy formula for estimating the reliability of a test of a given length, and the related formula for estimating the validity of a test of a given length, and it is implicit in the inferences usually made from item analyses and factor analyses of items. Systematic investigations of the generality of the effect found in this study and the amount of error that it introduces into these and other psychometric formulas and procedures seem necessary. If this effect is general and strong, it may be possible to modify the affected formulas, or, barring that, substitute empirical determinations of such characteristics as the optimum length of tests; this effect in item analyses and factor analyses of items can be appraised by including item location as a control variable in these procedures.

The findings about the relationships between the position of the correct alternative and the validity indexes are less consistent and correspondingly more difficult to interpret. These inconsistencies may be due to peculiarities in the samples or tests, or they may, perhaps, merely reflect sampling error, particularly since several of the correlations barely reach the .05 level of significance. Since all three tests were not administered to all three samples, however, it is not possible to tease apart the possible causes of these inconsistencies. The possibility that the regularities in the data are associated with the position of the alternative, per se, rather than its meaning, is suggested by the observation that in five of the six instances in which the correlations were significant, the items' correlations with either the content or response style score were higher when the second alternative was keyed correct. However, the way in which the position of the alternative would affect its correlation with the content or response style score is difficult to understand.

REFERENCES

- Banta, T. J. "Social Attitudes and Response Styles." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 543-557.
- Bass, B. M. "Authoritarianism or Acquiescence?" *Journal of Abnormal and Social Psychology*, LI (1955), 616-623.
- Bass, B. M. "Undiscriminated Operant Acquiescence." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVII (1957), 83-85.
- Berg, I. A. "The Reliability of Extreme Position Response Sets in Two Tests." *Journal of Psychology*, XXXVI (1953), 3-9.
- Brogden, H. E. "Implications of Item-Index Intercorrelations for Item Analysis." Technical Research Note 62. Washington: Department of the Army, Adjutant General's Office, 1956.
- Caldwell, E. "Stability of Scores on a Personality Inventory Administered during College Orientation Week." *Personnel Guidance Journal*, XXXVIII (1959), 305-308.
- Chall, Jeanne S. *Readability: An Appraisal of Research and Application*. Columbus, Ohio: Ohio State University Press, 1958.
- Couch, A. and Keniston, K. "Yeasayers and Naysayers: Agreeing Response Set as a Personality Variable." *Journal of Abnormal and Social Psychology*, LX (1960), 151-174.
- Cronbach, L. J. "Response Sets and Test Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 475-494.
- Cronbach, L. J. "Further Evidence on Response Sets and Test Design." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, X (1950), 3-31.
- Eisenberg, P. and Wesman, A. G. "Consistency in Response and Logical Interpretation of Psychoneurotic Inventory Items." *Journal of Educational Psychology*, XXXII (1941), 321-338.
- Fiske, D. W. "The Constraints on Intra-individual Variability in Test Responses." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVII (1957), 317-337.
- Fiske, D. W. "The Inherent Variability of Behavior." In D. W. Fiske and S. R. Maddi (Eds.), *Functions of Varied Experience*. Homewood, Ill.: Dorsey, 1961. Pp. 326-354.
- Fiske, D. W. and Rice, Laura. "Intra-individual Response Variability." *Psychological Bulletin*, LII (1955), 217-250.
- Frederiksen, N. "A Study of Tests of Report Writing Ability." ONR Technical Report. Princeton, N. J.: Educational Testing Service, 1958.
- Frederiksen, N. and Messick, S. "Response Set as a Measure of Personality." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIX (1959), 137-157.
- Gage, N. L., Leavitt, G. S., and Stone, G. C. "The Psychological Meaning of Acquiescence Set for Authoritarianism." *Journal of Abnormal and Social Psychology*, LV (1957), 98-103.
- Gordon, L. V. "The Effect of Position on the Preference Value of Personality Items." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XII (1952), 669-676.

- Hanley, C. "The 'Difficulty' of a Personality Inventory Item." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 577-584.
- Helmstadter, G. C. "Procedures for Obtaining Separate Set and Content Components of a Test Score." *Psychometrika*, XXII (1957), 381-393.
- Jackson, D. N. "Cognitive Energy Level, Acquiescence, and Authoritarianism." *Journal of Social Psychology*, LIX (1959), 65-69.
- Jackson, D. N. and Messick, S. "Content and Style in Personality Assessment." *Psychological Bulletin*, LV (1958), 243-252.
- Kaufman, P. "Changes in the Minnesota Multiphasic Personality Inventory as a Function of Psychiatric Therapy." *Journal of Consulting Psychology*, XIV (1950), 458-464.
- Lighthall, F. F., Davidson, K. S., Waite, R. R., Sarason, S. B., and Sarnoff, I. "The Effects of Serial Position and Time Interval on Two Anxiety Questionnaires." *Journal of General Psychology*, LXIII (1960), 113-131.
- Loevinger, Jane. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports*, III (1957), 635-694. (Monograph Supplement, 1957, No. 9-V3.)
- Marcus, A. "The Effect of Correct Response Location on the Difficulty Level of Multiple-Choice Questions." *Journal of Applied Psychology*, XLVII (1963), 48-51.
- McCreary, Joyce B. and Bendig, A. W. "Comparison of Two Forms of the Manifest Anxiety Scale." *Journal of Consulting Psychology*, XVIII (1954), 206.
- McGeoch, J. A. and Whitely, P. L. "The Reliability of the Pressey X-0 Tests for Investigating the Emotions." *Journal of Genetic Psychology*, XXXIV (1927), 255-270.
- Messick, S. "Separate Set and Content Scores for Personality and Attitude Scales." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 915-923.
- Mollenkopf, W. G. "An Experimental Study of the Effects on Item Analysis Data of Changing Item Placement and Test Time Limit." *Psychometrika*, XV (1950), 291-317.
- Rapaport, G. M. and Berg, I. A. "Response Sets in a Multiple-Choice Test." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XV (1955), 58-62.
- Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., and Ruebush, B. K. *Anxiety in Elementary School Children*. New York: Wiley, 1960.
- Stone, L. A. and James, R. L. "Stability of Agreeing Response Set over a Period of Time." *Psychological Reports*, VIII (1961), 350.
- Stricker, L. J. "Acquiescence and Social Desirability Response Styles, Item Characteristics, and Conformity." *Psychological Reports*, XII (1963), 319-341. (Monograph Supplement, 1963, No. 2-V12.)
- Thorndike, E. L. and Lorge, I. *The Teacher's Word Book of 30,000*

Words. N. Y.: Teachers College, Columbia University, Bureau of Publications, 1944.

Voas, R. B. "Comparison of the Taylor Anxiety Scale Administered Separately and Within the MMPI." *Psychological Reports*, II (1956), 373-376.

Whitcomb, M. A. and Travers, R. M. W. "A Study of Within-Test Learning Functions as a Determinant of Total Score." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVII (1957), 86-97.

Windle, C. "Test-Retest Effect on Personality Questionnaires." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIV (1954), 617-633.

Windle, C. "Further Studies of Test-Retest Effect on Personality Questionnaires." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XV (1955), 246-253.

AN ANALYSIS OF TEST-WISENESS¹

JASON MILLMAN AND CAROL H. BISHOP

Cornell University

ROBERT EBEL

Educational Testing Service²

THE purpose of this article is to offer an analysis of test-wiseness which can serve as a theoretical framework for empirical investigations. Thorndike (1949), Ebel and Damrim (1960), and Vernon (1962), among others, write that there may be sources of variance in educational test scores other than item content and random error. Test-wiseness was one suggested source.

General Definition "Test-wiseness" is defined as a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score. Test-wiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures. A somewhat limited concept of test-wiseness will be analyzed in this article. The analysis will exclude factors concerned with the general mental attitude (such as anxiety and confidence) and motivational state of the examinee, and it will be restricted to the actual taking of (not preparing for) *objective* achievement and aptitude tests.

Empirical Studies of Test-Wiseness There appears to be no systematic study of either the importance of test-wiseness or the degree to which it can be taught or measured. This is true even though both professional writers (e.g. Vernon, 1962; Anastasi, 1961; Pauck, 1950) and popular writers (e.g. Whyte, 1956; Huff, 1961; Hoff-

¹ This research was supported, in part, by the Cornell Social Science Research Center through a grant from the Ford Foundation.

² Now at Michigan State University.

mann, 1962) have referred to test-wiseness as a potentially large source of test variance. Almost all textbooks in educational measurement provide rules intended to help the test constructor avoid giving away the answer. Numerous pamphlets on how to take tests are available (e.g. Hook, 1958; Heston, 1953; and Pettit, 1960).

Although no comprehensive investigation of test-wiseness is known to exist, data from a few empirical studies indicate that it is a factor which may deserve attention.

Comparisons made by Vernon of British and American Students on different types of examinations "illustrate the importance of facility or sophistication at such tests" (1962, p. 285). This study concludes that "assessments of the understanding of complex concepts, whether by objective tests, written essays, oral or other methods, are affected not only by the level and type of concept tested, but also by many factors arising from the method of testing, and the subject's facility in handling it" (1962, p. 285).

In an unpublished study, Millman and Setijadi (in press) recently compared American and Indonesian students on a series of algebra items. The Indonesian students had never taken an objective test before, but during the practice session quickly learned the mechanics of reporting the best choice. The two groups scored equally well on the open-ended version of the items. Nevertheless, the performance of the American Students on the multiple-choice version (having plausible options) was significantly better than that of the Indonesians. The American students gained an additional advantage when the items contained one or more implausible options.

Test-wiseness seems to be responsible, in part, for the effects of practice and coaching on test scores. During the early twenties and thirties, a series of studies indicated that a substantial rise on the Stanford-Binet score was possible after practice and/or coaching with similar material. (See Rugg and Colloton, 1921, for review.) The effects of practice and coaching on the Moray House Test Examinations received great emphasis in England. The average gains from these sources were estimated at around nine I.Q. points, with greater gains among testees who were genuinely unsophisticated about tests to begin with (Vernon, 1954). The effects of practice and coaching on S.A.T. scores have not been so dramatic, perhaps because the students are more sophisticated (Levine and Angoff, 1958). Differences between coached and uncoached groups

are usually less than the standard error of the test, but they are in the positive direction and represent systematic bias.

Recent reviews of non-equivalence of measures of supposedly the same traits (Campbell and Fiske, 1959; Smith and Kendall, 1963) suggest the importance of the effects of testing method and situation on test scores. Test-wiseness may contribute appreciably to these effects.

There is some evidence that high school students can verbalize many principles of test-wiseness. Two hundred and forty high achieving students in a suburban high school were instructed by the senior investigator as follows.

Let us pretend a new student from a different part of the United States has just moved into this area and is now attending your school. He is having trouble getting good scores on the tests of some of the teachers. Where he comes from they do not give tests like the tests the teachers in this school give. What help can you be to this student? Write below any suggestions you can give the student. Tell him about things you have found successful when preparing for and taking tests in certain courses.

The students were assured their answers would be seen only by project personnel at Cornell.

Some of the responses elicited by this primarily *unstructured* questionnaire, together with the percent of students providing the response, are shown below:

General Response	Percent
Plan your time.	7
Answer easier questions first.	8
Do not spend too much time on one question (or come back later if you don't know the answer).	27
Read directions (or questions) carefully.	44
Recheck your answers for errors.	20
Guess if you don't know the answer.	18
Eliminate impossible foils.	17
Look for leads from other questions.	13
Don't read into questions (or answers) too deeply.	5
Watch for specific determiners.	3

Test-wiseness was cited as an important reason for success on examinations by college students also. Gaier asked 276 students to "assume that you will receive a letter grade of (A)/(D) on the test you are to take. List the specific activities, either on your part

or on the part of the instructor, that you feel were influential or responsible in making this grade" (1962, p. 561). Of the 136 students who were to assume they received a grade of A on the test, 21 percent volunteered "test understanding" as a reason for success, 21 percent indicated "comprehension and reasoning ability," and 18 percent suggested "test characteristics" as a reason for successful test performance. "Not able to understand and to reason" (26 percent of the responses), "not understanding tests" (13 percent), "teacher's tests" (12 percent), and "test characteristics" (34 percent) were volunteered by the 140 students who were to assume they received a grade of D on the test as reasons why their test performance was unsuccessful.

Further evidence of the importance of test-wiseness comes from investigations in which the problem solving styles of students answering objective type test items are studied (e.g., Bloom and Broder, 1950; Earle, 1950; Connerly and Wantman, 1964; French, 1965). Under the supervision of the authors, 40 college students in two institutions were interviewed individually as they were taking regular course examinations. (Time limits of the examinations were extended to permit the examinees to explain why they chose the answer they did at the time they responded to the item.) There was a great range in the sophistication of reasons given by students for responding as they did to test items in which they did not know the answers.

The particular styles reported are highly specific to the nature of the items being solved. However, the problem solving styles of high and low test performers could be distinguished. Bloom and Broder (1950) report that students trained in those general problem solving techniques (including the comprehension of test directions, the understanding of the nature of the specific test questions, and the ability to reason logically) used by high test performers, but *not* additionally trained in subject-matter knowledge, made significant gains in subsequent achievement test scores. French (1965) demonstrated that the factor composition of a test frequently depended upon the problem solving styles used in answering the test items.

The results of the references cited above and an analysis of the literature dealing with principles of test construction or advice for taking examinations were useful in developing the following outline.

An Outline of Test-Wiseness Principles

I. Elements independent of test constructor or test purpose.

A. Time-using strategy

1. Begin to work as rapidly as possible with reasonable assurance of accuracy.
2. Set up a schedule for progress through the test.
3. Omit or guess at items (see I.C. and II.B.) which resist a quick response.
4. Mark omitted items, or items which could use further consideration, to assure easy relocation.
5. Use time remaining after completion of the test to reconsider answers.

B. Error-avoidance strategy.

1. Pay careful attention to directions, determining clearly the nature of the task and the intended basis for response.
2. Pay careful attention to the items, determining clearly the nature of the question.
3. Ask examiner for clarification when necessary, if it is permitted.
4. Check all answers.

C. Guessing strategy.

1. Always guess if right answers only are scored.
2. Always guess if the correction for guessing is less severe than a "correction for guessing" formula that gives an expected score of zero for random responding.
3. Always guess even if the usual correction or a more severe penalty for guessing is employed, whenever elimination of options provides sufficient chance of profiting.

D. Deductive reasoning strategy.

1. Eliminate options which are known to be incorrect and choose from among the remaining options.
2. Choose neither or both of two options which imply the correctness of each other.
3. Choose neither or one (but not both) of two statements, one of which, if correct, would imply the incorrectness of the other.

4. Restrict choice to those options which encompass all of two or more given statements known to be correct.
5. Utilize relevant content information in other test items and options.

II. Elements dependent upon the test constructor or purpose.

A. Intent consideration strategy.

1. Interpret and answer questions in view of previous idiosyncratic emphases of the test constructor or in view of the test purpose.
2. Answer items as the test constructor intended.
3. Adopt the level of sophistication that is expected.
4. Consider the relevance of specific detail.

B. Cue-using strategy.

1. Recognize and make use of any consistent idiosyncracies of the test constructor which distinguish the correct answer from incorrect options.
 - a. He makes it longer (shorter) than the incorrect options.
 - b. He qualifies it more carefully, or makes it represent a higher degree of generalization.
 - c. He includes more false (true) statements.
 - d. He places it in certain physical positions among the options (such as in the middle).
 - e. He places it in a certain logical position among an ordered set of options (such as the middle of the sequence).
 - f. He includes (does not include) it among similar statements, or makes (does not make) it one of a pair of diametrically opposite statements.
 - g. He composes (does not compose) it of familiar or stereotyped phraseology.
 - h. He does not make it grammatically inconsistent with the stem.
2. Consider the relevancy of specific detail when answering a given item.
3. Recognize and make use of specific determiners.
4. Recognize and make use of resemblances between the options and an aspect of the stem.

5. Consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.

General Description of the Outline The suggested outline is divided into two main categories: elements which are independent of the test constructor or test purpose and those which are dependent upon the constructor or purpose. The principles falling into the former category are potentially profitable regardless of previous contact with the test constructor or previous contact with tests having a similar purpose.

The first two subdivisions, time-using and error-avoidance strategies, contain strategies which allow the examinee to demonstrate fully his knowledge of the specific subject matter by helping him avoid *losing* score points for reasons other than lack of knowledge of the test content.

Time-using strategy applies only to those tests which restrict the time allotted the examinee. In general, the elements are concerned with the most efficient use of allotted time. Error-avoidance strategy applies to all testing situations. The elements are simply concerned with the avoidance of careless mistakes.

The last two subdivisions in the first category, guessing and deductive reasoning strategies, contain strategies which allow the examinee to gain points beyond that which he would receive on the basis of sure knowledge of the specific subject matter.

Guessing strategy may allow the examinee to receive credits for responses made on a completely chance basis; deductive reasoning strategy deals with methods of obtaining the correct answer indirectly or with only part of the knowledge necessary to answer a question. It should be noted that *some* knowledge of the subject matter is involved in deductive reasoning strategies, but the correct answer itself would not be known if no choices were given or no other questions were asked.

The second main category, elements dependent upon the test constructor or purpose, includes those strategies which the examinee may employ only after knowing the test constructor's views or the test purpose, or after having had contact with and feedback from similar tests.

The first subdivision in this category is consideration of intent. It is similar to the first two subdivisions in the previous category in that it is concerned with strategies which allow the examinee to

avoid being penalized for anything other than lack of knowledge of the subject matter of the test.

The second subdivision, cue-using strategy, pertains to the use of cues which are available when a specific answer is not known. As in the last two subdivisions in the previous category, partial knowledge of the subject matter may be needed. The cues can be used successfully, however, only to the extent that a correlation has been established, under similar conditions, between the cues and the correct answer.

Further description of selected elements, with examples, may clarify the tactics and strategies.

Elaboration of Selected Principles

I.A.1. Begin to work as rapidly as possible with reasonable assurance of accuracy.

The pace at which one can work without sacrificing accuracy varies with individuals. One should attempt, however, to complete the test in less time than is allotted in order to allow time to check answers.

I.A.2. Set up a schedule for progress through the test.

A rule of thumb is to determine how far one should be when a specific proportion of the testing period has elapsed. A periodic check on rate of progress facilitates the maintenance of proper speed (Cook, 1957). This principle suggests the necessity of determining the scope of the test before beginning work.

I.A.3. Omit or guess at items which resist a quick response.

When time is limited the examinee should work first on those items which will yield the most points in a given amount of time. The order in which he works on the items may be determined by the relative difficulty of items, by the relative time needed to read and answer the items, or by possible heavier weighting of some items. If such differences are not apparent, the examinee should work in order of presentation of the items.

I.A.5. Use time remaining after completion of the test to reconsider answers.

Change answers if it seems desirable. Examinees generally increase their scores when they do (Berrien, 1939, Briggs and Reile, 1952). There is some evidence that persistence (i.e. using full time on test) pays off (Briggs and Johnson, 1942).

I.B.2. Pay careful attention to the items, determining clearly the nature of the question.

Guard against inferring the answer before completely reading the question. Exercise special care in answering more complex questions such as negatively stated items and items having more than one clause.

I.C. Guessing strategy

A distinction can be made between informed and blind guessing. This subdivision is concerned with blind guessing, i.e. selecting the answer at random without considering the content of the options. Although it is recognized that students rarely respond to items in this way and that they should be encouraged to make informed rather than blind guesses, there are times when, because of lack of time or motivation, blind guessing is used. These strategies may also be of value to the examinee who, on the basis of the item content, is able to assign to the options probabilities of being correct. (But see caution in II. B.)

The validity of any set of recommended strategies for guessing depends upon what the examinee is attempting to maximize. The strategies listed in subdivision I.C. are based upon the assumption that the examinee wishes to maximize the expected value of his test score.³ This expected value is a function of the probability that the correct option is selected and the utilities associated with correct and incorrect answers. This relation may be expressed algebraically as:

$$E.V. = P_c U_c + (1 - P_c) U_i$$

where E. V. is the expected value, P_c is the probability that the correct option is selected and U_c and U_i are the utilities associated with correct and incorrect choices.

I.C.1. Always guess if right answers only are scored.

Example:

Consider a four alternative multiple-option item in which one point is given for the correct answer and no points are subtracted

³ If a minimax decision function were used, that is, an examinee wanted to minimize his maximum loss, he would never guess when a penalty for guessing was employed. As another example, suppose the testee wished to guess only when the probability of improving his score by guessing was greater than one-half. If the usual correction for guessing formula was employed, he would guess only when $(n+1)/k = \text{an integer } (k > 2)$, where n is the number of items to be guessed and k the number of choices per item. (Graesser, 1958).

for an incorrect answer (i.e. scored rights only). A person who makes a pure guess is expected to earn

$$E.V. = 1/4 (1) + (1 - 1/4) (0) = 1/4 \text{ point.}$$

Since $1/4$ is greater than 0, the value of an omitted item, the examinee should guess.

- I.C.2. Always guess if the correction for guessing is less severe than a "correction for guessing" formula that gives an expected score of zero for random responding.

Example:

If, in the above illustration, $1/3$ of a point were subtracted for an incorrect answer, then:

$$E.V. = 1/4 (1) + (1 - 1/4) (-1/4) = 1/16 \text{ point.}$$

This value is greater than that of an omitted item, and therefore, the examinee should guess.

- I.C.3. Always guess, even if the usual correction or a more severe penalty for guessing is employed, whenever elimination of options provides sufficient chance of profiting.

Example:

Assume that in the above illustration, the usual correction for guessing formula were employed, and thus $1/3$ of a point is deducted for an incorrect answer. Assume further that the examinee can definitely eliminate one incorrect option, that is, there is no guessing involved in eliminating this option. The expected value associated with choosing from among the remaining three options is:

$$E.V. = 1/3 (1) + (1 - 1/3) (-1/3) = 1/9 \text{ point.}$$

Again, the examinee should guess.

- I.D. Deductive reasoning strategy.

The test-wise person who does not know the correct option directly may be able to deduce the answer by logical analysis or by using information gained from other items. This strategy differs from cue-using strategy in that it is not necessary to establish correlations between cues and the correct answer. The following strategies may be used successfully in any objective testing situation, their success depending upon the examinee's ability to reason in a logically valid manner.

- I.D.1. Eliminate options which are known to be incorrect and choose from among remaining options.

The examinee may be able to eliminate some options with par-

tial knowledge of the subject matter. Options may often be eliminated because they are logically inconsistent with the stem.

Examples:

The state with the highest total population in the United States in 1950 was:

- *a) New York
- b) Chicago
- c) Michigan
- d) California

Option b is inconsistent with the stem since it is not a state, and the choice is, therefore, restricted to a, c, or d.

Which one of the following is an advantage of using high beds in hospitals?

- a) High beds cost more than regular size beds.
- *b) The care of patients is less difficult when the beds are high.
- c) High beds can be used all day.
- d) People are less likely to fall out of high beds.

Since high cost is never an advantage to the buyer, option a is logically inconsistent with the stem and may be eliminated.

I.D.2. Choose neither or both of two options which imply the correctness of each other.

Example:

A mental disorder which is often classified as a neurosis is:

- *a) hysteria
- b) dementia praecox
- c) schizophrenia
- d) involuntional melancholia

If only one answer can be chosen and the examinee knows that "dementia praecox" is another name for "schizophrenia," he knows the answer must be either a or d. Care must be taken in deciding whether two options do, in fact, imply the correctness of each other.

I.D.3. Choose neither or one (but not both) of two statements, one which, if correct, would imply the incorrectness of the other.

This situation occurs most frequently with items containing options which are the opposite of each other.

Example:

The 18th amendment to the U. S. Constitution

- a) prohibited the manufacture, sale or transportation of intoxicating liquors within the United States.

- b) repealed the prohibition amendment.
- c) gave women the right to vote.
- d) prohibited the President from being elected to office more than twice.

Because the correctness of option a implies the incorrectness of option b, both option a and option b cannot be correct. Note, however, that had the item stem asked about the nineteenth amendment, neither option a nor option b would have been correct. (See II.B.1.f.)

I.D.4. Restrict choice to those options which encompass all of two or more given statements known to be correct.

Examples:

Which of the following men were presidents of the United States?

- a) George Washington
- b) Andrew Jackson
- c) Abraham Lincoln
- *d) All of the above

A test-wise examinee who was sure that Washington and Lincoln were presidents but was undecided about Jackson would nevertheless select option d.

A statistical average is a:

- a) mean
- b) mode
- c) median
- *d) measure of central tendency

The more inclusive option d would be chosen by a test-wise person who knew that at least two of the first three choices were correct and that there was only one keyed answer.

I.D.5. Utilize relevant content information in other test items and options.

Examples:

Which one of the following four animals is warm-blooded?

- a) snake
- b) frog
- *c) bird
- d) lizard

Which one of the following four animals is cold-blooded?

- *a) snake
- b) dog

c) kangaroo

d) whale

Assume the examinee knows that a bird is a warm-blooded animal, and that all animals are either warm-blooded or cold-blooded but not both. He can then reason that a snake, which is an option in both items, must be a cold-blooded animal and can, therefore, answer the second question.

II.A. Intent consideration strategy.

It is possible that the examinee may receive a low score in an objective test merely because his views differ from the test constructor's or his level of knowledge is higher than that being tested. By recognizing and acting upon the biases of the test constructor or the intent of the test, the examinee may avoid loss of points due to misinterpretation, rather than lack of knowledge of the subject matter. (The use of this strategy assumes that the goal of the examinee is to earn the greatest possible number of points; i.e., he is not willing to lose points on principle.)

II.A.1. Interpret and answer questions in view of previous emphasis of the test constructor or in view of the test purpose.

It occasionally is difficult to determine what a question is "getting at," and as a consequence the answer is elusive. If the test is taken with a set for the test constructor's views or purposes, one may be able to interpret the question more easily.

Example:

There is a test for normality. T F

If the examinee knows that the purpose of the test is to measure statistical knowledge rather than knowledge of abnormal psychology, the question may be interpreted easily.

II.A.2. Answer items as the test constructor intended.

If the examinee believes that the test constructor had a certain answer in mind, but the examinee can think of a possible objection to the answer, he should answer the item as he believes the test constructor intended. That is, he should choose the option he believes has the greatest chance of being correct, even though others have merits and even though the chosen option is not completely satisfactory. This assumes that no explanation can be communicated to the grader; only the answer may be marked.)

Example:

Thomas Jefferson wrote the Declaration of Independence. T F

The examinee may object to answering "true" on the basis that the statement is not complete since four other men were also appointed to the committee in charge of writing it. The test-wise person, however, would answer "true" if he felt this was the answer which the test constructor considered correct.

II.A.3. Adopt the level of sophistication that is expected.

An item (especially a true-false item) may have different correct answers depending upon the "depth" which the constructor or purpose demands. The test-wise person will recognize and adopt the appropriate level.

Example:

Light travels in a straight line. T F

This item may be keyed as true in a test given at an elementary grade level, but may be false in an advanced college course in physics.

II.A.4. Consider the relevance of specific detail.

Depending upon the test constructor or test purpose, the specific detail in an item may or may not have bearing upon the answer.

Example:

A picture of George Washington, the most influential man in shaping the history of the United States, appears on the one dollar bill. T F

The test-wise person would have to decide whether the appositive were a mere insertion by an enthusiastic admirer of Washington, or whether it had relevance to the answer.

II.B. Cue-using strategy

Since one test constructor may inadvertently give away the correct answer, whereas another may use these same cues as foils, the successful use of cues is dependent upon previous contact with similar tests to establish the relationship between the cues and the correct answer.

These cues should be used only when the examinee is unable to arrive at the answer using his knowledge of the subject matter and his reasoning ability.

Once the examinee has detected a cue, he is faced with the decision whether or not to make an *informed* guess. The blind guessing strategies may be of help here, with the replacement of subjective for objective probabilities in the expected value formulas. In assigning subjective probabilities, the examinee should bear in mind that

The item writer attempts to make each wrong response so plausible that *every* examinee who does not possess the desired skill or ability will select a wrong response. . . . In actual practice, this aim of the item writer is never fully realized, but it is doubtless often sufficiently realized that the standard formula markedly *over-corrects* (Traxler, 1951; p. 349).

There is some empirical evidence that this is actually the case (e.g. Gupta and Penfold, 1961; Little, 1962).

II.B.1. Recognize and make use of any consistent idiosyncracies of the test constructor which distinguish the correct answer from incorrect options.

II.B.1.e. He places it in a certain logical position among an ordered set of options (such as the middle of the sequence).

Example:

The population of Albany, New York in 1960 was approximately

- *a) 130,000
- b) 460,000
- c) 75,000
- d) 220,000

To guard against the examinee who greatly over- or underestimates the desired value answering the item correctly, the inexperienced test maker usually favors including foils with values more extreme (in both directions) than the value of the correct option.

II.B.1.f. He includes (does not include) it among similar statements, or makes (does not make) it one of a pair of diametrically opposite statements.

Examples containing similar statements:

In a second class lever the

- a) effort is between the fulcrum and the weight.
- *b) weight is between the fulcrum and effort.
- c) fulcrum is not used.
- d) mechanical advantage is less than one.

Behavior for which the specific eliciting stimulus is not determined is called

- *a) operant
- b) operational
- c) apopathic
- d) prehensory

Example containing opposite statements:

Adding more items to a test would probably

- a) decrease its reliability
- *b) increase its reliability
- c) decrease its cost of administration
- d) increase its standard error of measurement

II.B.1.g. He composes (does not compose) it of familiar or stereotyped phraseology.

Example:

Behavior for which the specific eliciting stimulus is not determined is called

- a) apopathic
- *b) operant
- c) abient
- d) prehensory

If "operant" is the only word the student recalls hearing or reading, he may tend to select it, even though its meaning is unclear.

II.B.2. Consider the relevance of specific detail when answering a given item.

Examples:

According to your textbook, the best way to teach children an activity is to have them perform the activity themselves. T F

The Iliad, written by Homer, is an early Greek epic. T F

In the above examples, if the test-wise person did not know whether the textbook made the specific statement or whether Homer wrote the Iliad (but did know the truth of the main statement) he would need to determine whether these details were a mere insertion of no consequence or whether such details usually had relevance to the answer.

The distinction between this principle and principle II.A.4. is whether or not the specific detail in a test item is known to be true by the examinee.

II.B.3. Recognize and make use of specific determiners.

The following lists contain examples of words which may be correlated with correct or incorrect answers respectively:

often	always
seldom	never
perhaps	necessarily
sometimes	only
generally	merely
may	must
etc.	etc.

II.B.4. Recognize and make use of resemblances between the options and an aspect of the stem.

These resemblances may take the form of a direct repetition, synonym or more general associative connection.

Example of direct repetition:

The aeronautics board which has jurisdiction over civil aircraft is called:

- *a) Civil Aeronautics Board
- b) Committee to Investigate Airlines
- c) Division of Passenger Airways

Examples of general associative connections:

"Glean

- 1. polish
- 2. gather
- 3. skim
- 4. praise" (Votaw, 1955; Form B, p. 4)

The similarity between the sounds and spelling of glean and gleam creates an associative connection which would lead the test-wise person to consider option 1 as merely an attractive foil.

"Descriptions

Titles

- (E) 1. Planting a sloping field alternately with rows of corn, then rows of wheat, then rows of corn, etc.
- (D) 2. Plowing a crop underground instead of harvesting it.
- (A) 3. Removing brush and weeds along the fence between the fields.
- (B) 4. Planting around a hillside in level rows instead of planting up and down over the hill.
- (C) 5. Planting a field one year with wheat, the second year with oats, the third year with alfalfa, the fourth year with corn.

A. Clean farming

B. Contour farming

C. Crop rotation

D. Green manuring

E. Strip cropping"
(Ahmann and Glock,
1963; p. 101)

Associations such as alternate rows—strip; removing—clean; hillside rows—contour; and one year wheat, second year oats—rotation lessen the difficulty of the item for the test-wise person.

II.B.5. Consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.

A question may be embedded among questions dealing with the same general subject, and the given context may make it possible to deduce the answer.

Examples:

The 13th, 14th, and 15 amendments to the Constitution are commonly called the:

- *a) reconstruction amendments
- b) Big Three amendments
- c) Bill of Rights
- d) presidential amendments

If the item were embedded among items dealing with the post Civil War period, the test-wise person would be given a cue to the correct answer.

"What is the chief obstacle to effective homogeneous grouping of pupils on the basis of their educational ability?"

- a) Resistance of children and parents to discriminations on the basis of ability.
- b) Difficulty of developing suitably different teaching techniques for the various levels.
- c) Increase costs of instruction as the number of groups increases and their average size decreases.
- *d) Wide differences in the level of development of various abilities within individual pupils." (National Council on Measurement in Education, 1962; p. 239)

If this item were embedded among items dealing with educational testing and evaluation, the test-wise person would favor option d. Resistance of children, teaching techniques, and costs of instruction are not of the same species as the questions and options of other items in the test.

Implications for Testing. The analysis of test-wiseness proposed in this paper is intended to serve as a framework to study its importance. If it does make a significant difference, it would be desirable to seek ways to reduce differences in test-wiseness among examinees in order to provide more valid estimates of their actual abilities and achievement levels. Examples of specific questions, which would then be of interest, follow. How can we change test items and test directions, or other conditions of administration, to minimize harmful effects of differences in test-wiseness? How can we more nearly equalize test-wiseness among examinees? Can it be taught? If so, how long will it take and can guide lines be published? What are the correlates of test-wiseness? Where is it in the spectra of intelligence? Is it related to the sheer number of tests taken? Does knowledge of how tests are constructed increase test-

wiseness? Is the degree of test-wiseness of an individual dependent upon the subject matter of the test? At what grade level does evidence of the various aspects of test-wiseness first appear? To help answer these questions, valid measures of test-wiseness are desired.

REFERENCES

- Ahmann, J. Stanley and Glock, Marvin D. *Evaluating Pupil Growth*. Boston: Allyn and Bacon, Inc., 1963.
- Anastasi, Anne. *Psychological Testing*. New York: The Macmillan Company, 1961.
- Berrien, F. K. "Are First Impressions Best on Objective Tests?" *School and Society*, L. (1939), 319-20.
- Bloom, Benjamin S. and Broder, Lois J. *Problem-solving Processes of College Students: An Exploratory Investigation*. Chicago: The University of Chicago Press, 1950.
- Briggs, Arvella and Johnson, D. M. "A Note on the Relation between Persistence and Achievement on the Final Examination." *Journal of Educational Psychology*, XXXIII (1942), 623-27.
- Briggs, Leslie J. and Reile, Patricia J. "Should Students Change Their Initial Answers on Objective-Type Tests?: More Evidence Regarding an Old Problem." *Journal of Educational Psychology*, XLIII (1952), 110-15.
- Campbell, Donald T. and Fiske, Donald W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin*, LVI (1959), 81-105.
- Connerly, John A. and Wantman, Morey J. "An Exploration of Oral Reasoning Processes in Responding to Objective Test Items." *Journal of Educational Measurement*, I (1964), 59-64.
- Cook, Desmond. "A Comparison of Reading Comprehension Scores Obtained Before and After a Time Announcement." *Journal of Educational Psychology*, XLVIII (1957), 440-46.
- Earle, Dotsie. "A Study of Problem-Solving Methods Used on Comprehensive Examinations." Chicago: University Examiner's Office, University of Chicago, 1950. (Mimeographed)
- Ebel, Robert L. and Damrin, Dora E. "Tests and Examinations." In Chester Harris (Editor) *Encyclopedia of Educational Research*. New York: The Macmillan Company, 1960. (Pp. 1502-517.)
- French, John W. "The Relationship of Problem-Solving Styles to the Factor Composition of Tests." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXV (1965), 9-28.
- Gaier, Eugene L. "Students' Perceptions of Factors Affecting Test Performance." *Journal of Educational Research*, LV (1962), 561-66.
- Graesser, R. F. "Guessing on Multiple-Choice Tests." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVIII (1958), 617-20.
- Gupta, R. K. and Penfold, D. M. Edwards. "Correction for Guessing in True-False Tests: An Experimental Approach." *British Journal of Educational Psychology*, XXXI (1961), 249-56.

- Heston, Joseph C. *How to Take a Test*. Chicago: Science Research Associates, Inc., 1953.
- Hoffmann, Banesh. *The Tyranny of Testing*. New York: Crowell-Collier Press, Inc., 1962.
- Hook, J. N. *How to Take Examinations in College*. New York: Barnes & Noble, Inc., 1958.
- Huff, Darrell. *Score: The Strategy of Taking Tests*. New York: Appleton-Century-Crofts, Inc., 1961.
- Levine, Richard S. and Angoff, William H. "The Effects of Practice and Growth on Scores on the Scholastic Aptitude Test." Research and Development Report No. 58-6/SR-586. Princeton, N. J.: Educational Testing Service, 1961.
- Little, E. B. "Overcorrection for Guessing in Multiple-Choice Test Scoring." *Journal of Educational Research*, LV (1962), 245-52.
- Millman, Jason and Setijadi. "A Comparison of American and Indonesian Students on Three Types of Test Items." *Journal of Educational Research*, in press.
- National Council on Measurement in Education. (Wilbur L. Layton, Secretary-Treasurer). "Multiple-Choice Items for a Test of Teacher Competence in Educational Measurement." Ames, Iowa: Iowa State University, 1962.
- Pauck, Charles E. "A Square Deal for the Examination." *Education*, LXXI (1950), 222-25.
- Pettit, Lincoln. *How to Study and Take Exams*. New York: John F. Rider Publisher, Inc., 1960.
- Rugg, Harold and Colloton, Cecile. "Constancy of the Stanford-Binet I.Q. as Shown by Retests." *Journal of Educational Psychology*, XII (1921), 315-22.
- Smith, Patricia Cain and Kendall, Lorne M. "Cornell Studies of Job Satisfaction: VI Implications for the Future." Cornell Notes in Psychology. Ithaca, New York: 1963. (Mimeographed)
- Thorndike, Robert L. *Personnel Selection: Test and Measurement Techniques*. New York: John Wiley & Sons, 1949.
- Traxler, Arthur E. "Administering and Scoring the Objective Test." In E. F. Lindquist (Editor), *Educational Measurement*. Washington: American Council on Education, 1951. (Chapter X.)
- Vernon, Philip E. "Symposium on the Effects of Coaching and Practice in Intelligence Tests: Conclusions." *British Journal of Educational Psychology*, XXIV (1954, Part 2), 57-63.
- Vernon, Philip E. "The Determinants of Reading Comprehension." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 269-86.
- Votaw, David F. *High School Fundamentals Evaluation Test*. Austin, Texas: The Steck Company, 1955.
- Whyte, William H., Jr. *The Organization Man*. New York: Simon and Schuster, 1956.

Recent Reference

- Gibb, Bernard Gordon. "Test-wiseness as Secondary Cue Response." Unpublished doctoral dissertation, Stanford University, 1964.

THE EFFECTS OF GUESSING ON THE STANDARD ERROR OF MEASUREMENT AND THE RELIABILITY OF TEST SCORES

DALE MATTSON

University of Washington

IN recent articles Lord (1957, 1959) proposed the use of the binomial distribution as a means of estimating the standard error of measurement of test scores. He states that the conditions under which this method of estimating S_e is appropriate are as follows:

1. The test items are scored 0 or 1.
2. The score is the sum of the item scores.
3. The response to an item depends only on the item itself.
4. The "standard error of measurement" is defined as the standard deviation of the scores that an examinee would be expected to obtain on a large number of parallel test forms where the individual remains unchanged during testing.

Under these conditions the standard error of measurement is defined as follows:

$$S_e = \sqrt{NP_o(1 - P_o)} \quad (1)$$

where

N = number of items on the test,

P_o = percentage of items which the person would get correct in the entire population of possible items.

A true score would be defined as follows:

$$T = NP_o. \quad (2)$$

It is important to notice that P_o is defined as percentage correct

and not as percentage known (P_k). When guessing is possible the formula for P_o would be:

$$P_o = P_k + P_g, \quad (3)$$

where P_g = percentage guessed correctly in entire population.
When all students attempt all the questions the formula becomes:

$$P_o = P_k(1 - \pi) + \pi, \quad (4)$$

where π = probability of a correct guess.

Guessing and S_o

The use of formula (1) for S_o indicates that for tests of the same length any two examinees with the same true score would have the same S_o . This would be true even though one score represents items known plus items guessed and another represents only the items known. From this it would seem that guessing does not affect the S_o . Actually guessing does have an effect on S_o since guessing effects P_o . In fact guessing may either increase or decrease S_o !

Suppose that an examinee with $P_k = .80$ chooses not to guess at any items he does not know. On two-alternative, multiple-choice tests of 100 items his true score would be 80 with S_o equal to 4. Suppose also that the tests are of such a nature that π equals the reciprocal of the number of alternatives. Under these conditions if the examinee decided to guess, his true score would be 90 with S_o equal to 3.

In general, if the effect of guessing is to make P_o closer to .50, S_o will be increased. When guessing makes P_o deviate further from .50, S_o will be decreased.

Guessing and Reliability

As might be expected guessing reduces reliability even when S_o is reduced. This is because reliability is determined on the basis of the relationship between error variance (individual variation about true scores) and the true variance (group variation in true scores).

$$\begin{aligned} \text{and} \quad r_{11} &= 1 - \frac{S_o^2}{S_e^2}, \\ S_e^2 &= S_i^2 + S_o^2, \\ \therefore r_{11} &= 1 - \frac{S_o^2}{S_i^2 + S_o^2}. \end{aligned} \quad (5)$$

An examination of equation (5) reveals that a decrease in S_e^2 when S_i^2 remains constant results in an increase in reliability. In order for a decrease in S_e^2 to reduce reliability there must also be a reduction in S_i^2 . Therefore if guessing is to reduce both reliability and S_e^2 , guessing must also reduce S_i^2 .

From equations (2) and (4),

$$\text{and} \quad T = N(1 - \pi)P_k + \pi N,$$

$$S_i^2 = N^2(1 - \pi)^2 S_{P_k}^2. \quad (6)$$

From equation (6) it is evident that for a constant value of $S_{P_k}^2$ as Π (probability of a correct guess) increases, the variance of the true scores decreases. When Π equals zero $S_i^2 = N^2 S_{P_k}^2$.

Formula (5) can be rewritten as follows:

$$S_i^2 = \frac{r_{ii} S_e^2}{1 - r_{ii}},$$

or

$$S_i^2 = \frac{r_{ii} N P_c (1 - P_c)}{1 - r_{ii}}, \quad (7)$$

where P_c = average P_c for the group tested.
Then formulas (6) and (7) yield the following equality.

$$\frac{r_{ii} P_c (1 - P_c)}{1 - r_{ii}} = N(1 - \pi)^2 S_{P_k}^2. \quad (8)$$

By use of the formula (8) it is possible to determine how guessing affects the reliability of a test. Suppose that a recall test of 100 items is given where the average P_c for the group equals .50 and $r_{ii} = .80$. By the use of formula (8) with $\pi = 0$ the variance of P_k is found to be .01. Now suppose the test is made into a two-alternative, multiple-choice test where P_k is not changed but π becomes .50. Through use of equation (4) P_c is found to be .75. Then by substitution in equation (8), r_{ii} is found to be .5714.

The procedure which has just been described was used to obtain the information presented in Table 1. This table gives the expected reliabilities when recall tests with average P_k equal to .50, are changed to multiple-choice form. For example if a recall test with a reliability of .70 and an average P_k of .50 were changed to a five-alternative, multiple-choice test, the expected reliability would be .608. This, of course, assumes that all five of the alternatives are equally attractive to an examinee who does not know the correct

TABLE 1

Expected Changes in the Reliabilities of Tests Due to Changes in the Number of Alternatives: $P_k = .50$

Number of Alternatives	Reliabilities				
∞ (recall test)	.900	.800	.700	.600	.500
5	.857	.727	.608	.500	.400
4	.844	.706	.583	.474	.374
3	.815	.662	.553	.424	.329
2	.750	.571	.437	.330	.250

answer. Similarly if the test were made into a two-alternative test, the expected reliability would be .437. The information in Table 1 isn't restricted to tests of items numbering 100 but would be applicable to tests of any lengths.

Conclusions

The formulas in this paper are only applicable under conditions which seldom if ever would be met in actual practice. It is probably impossible in any testing situation to construct multiple-choice tests in such a way that π is the reciprocal of the number of alternatives. It is also doubtful that S_{P_k} would remain constant when a recall test is changed to multiple-choice form. The formulas given here do present a model, however, which serves to clarify the relationship of guessing to the standard error of measurement and reliability. Empirical evidence could be obtained to determine how nearly the real world conforms to this theoretical model. This evidence could be obtained by administering a well constructed standardized test to similar groups using forms which have been modified as to the number of alternatives.

REFERENCES

- Lord, F. M. "Do Tests of the Same Length Have the Same Standard Errors of Measurement?" *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVII (1957), 510-521.
- Lord, F. M. "Tests of the Same Length Do Have the Same Standard Errors of Measurement." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIX (1959), 233-239.

THE INTERACTION OF PERSONALITY AND INTELLIGENCE IN TASK PERFORMANCE

DAVID KIPNIS AND CARL WAGNER^{1, 2}

Naval Medical Research Institute
Bethesda, Maryland

PERSONALITY measurement has been traditionally concerned with the identification and measurement of independent dimensions of personality. More recently there has been a convergence of interests from several areas of psychology on the specification of conditions under which a given personality trait may be related to some criterion behavior. Thus motivational theorists (Taylor and Spence, 1952; Mandler and Sarason, 1952) have been concerned with conditions under which personality traits, conceived of as drive states, may facilitate or impede task performance. Social psychologists have been concerned with differing relationships between personality and performance as a function of leadership styles (Haythorn, et al., 1956; Fiedler, 1962), of differences in the social setting (Meyers, 1962; McGrath, 1962), or of differences in the incentive values and difficulty of the task (Atkinson, 1957; Feather, 1963). Finally, the introduction of the concept of moderator variable by Saunders (1956) has stimulated considerable interest in the detection of trait interactions within an individual which modify the relationship between a given test variable and subsequent behavior.

¹ The opinions and statements contained herein are the private ones of the writers and are not to be construed as official or reflecting the views of the Navy Department or the Naval Service at large.

² We wish to express appreciation to Commander A. D. Garvin, USN, for providing men from his command to serve as subjects. We also wish to acknowledge the help of Charles A. Morris and Pearl Thomas, who served as technical assistants during all phases of the study.

This paper reports a laboratory experiment concerned with the question of conditions under which certain personality variables are related to task performance. The personality variables were initially developed as part of a research program to identify noncognitive measures predictive of school and job performance of Navy enlisted men (Kipnis and Glickman, 1962). In several studies which attempted to predict these criteria, it was found that the validities of two tests seemed to be moderated by the general intelligence level of the recruits (Kipnis, 1962, 1965). The first test, a purported measure of persistence beyond minimum standards called the Hand Skills Test, most frequently predicted among lower ability men, but either did not predict or had negative validity among higher ability men. The second test, a purported measure of psychopathy, called the Insolence Scale, was found to be negatively related to performance among higher ability men, but had no predictive validity, or in some instances predicted with opposite sign, among lower ability men. Since up to two years elapsed between testing of the Ss and collection of criterion information, the significant predictions obtained suggest that the tests are measuring meaningful components of behavior. To date, the Insolence Scale and the Hand Skills Test have not been found to be significantly correlated with each other, or with measures of general intelligence.

In a previous article two possible explanations for the interaction effects with intelligence were proposed (Kipnis, 1965). On the one hand it was felt the results may represent a genuine interaction with cognitive processes. That is, it can be argued that the more intelligent employ different cognitive processes from those employed by the less intelligent (Osler and Trautman, 1961). Somehow then, such differences in cognitions interact with personality to produce the differential relationships found in the previous studies. Alternately, the results may represent an interaction between personality and task difficulty. Within the context of this explanation, the function of intelligence could be viewed as defining how difficult the work will be for the individual to grasp. To the extent that the work was relatively easy for higher ability men, the possibility exists that persons high in insolence became rapidly bored with their work. Conversely lower ability men probably had to keep "plugging" away to maintain an average level of task performance and the interest of high insolent individuals may have been sustained

by the challenge of the work. The reverse of this argument may be applied to results obtained with the persistence test. That is, as the work became difficult for some persons because of their relatively low abilities, persistence helped elevate their levels of performance.

The planning of this study was guided in large part by an effort to evaluate each of these two alternatives. This was done by having men scoring high and low on the Insolence Scale, and men scoring high and low on the Hand Skills Test, work on a series of tasks, each of which varied in task difficulty. To take into account possible interactions with intelligence, Ss were also grouped in terms of their intelligence level. Further, several studies have suggested the importance of the variables of task attraction and task expectation in relationship to personality variables (Feather, 1963; Atkinson, 1957). Accordingly, measures of these variables were obtained from all Ss at various times during the experiment and related to both task performance and the two personality tests.

Procedure

Subjects. Experimental Ss were men awaiting the commencement of training at Navy trade schools located at Bainbridge, Maryland. A total of 140 Ss were used in groups of approximately 14 to 16 men per day. At the time of the study these men had been in the Navy for approximately four months and ranged in age from 17 to 20 years.

Experimental Tasks. The following two tasks selected to represent a cognitive task and a motor tracking task were used.³

a. *Anagrams.* Anagrams were prepared at three levels of task difficulty, following procedures outlined by Mayzner and Tresselt (1958). Easy anagrams consisted of four letter words rated in the Thorndike-Lorge Teachers Dictionary as being Very Frequently or Frequently used in the general population. Only adjacent letters were transposed. Moderately difficult anagrams consisted of five letter words rated as Very Frequently or Frequently used, with no two of the original letters adjacent to each other. Difficult anagrams consisted of five letter words rated as being Infrequently used, with no two of the original letters adjacent to each other.

³ A reaction time task was also included. Because of time limitations, only half of all Ss completed this task. In no instance was performance with this measure related to the two personality tests.

The anagrams were arranged in booklet form, with 15 anagrams placed upon a page. The first eight pages contained easy anagrams; the next four moderately difficult anagrams; and the last four difficult anagrams. Performance was timed with one and one-half minutes per page allowed for the easy anagrams, six minutes per page allowed for the moderately difficult anagrams, and ten minutes per page allowed for the difficult anagrams. Average number of easy anagrams solved per minute was 4.76; average number of moderately difficult anagrams solved per minute was .69; and average number of difficult anagrams solved per minute was .34.

Ss were told that they were to solve as many anagrams as possible in the time allowed and not to spend too much time on any problem. They were each given two practice problems to solve using easy anagrams.

b. Pursuit Rotor. A Lafayette Pursuit Rotor with four different speeds of rotation (15 RPM; 30 RPM; 45 RPM; 60 RPM), was used. Time off-target (error scores), measured in 15th of a second, was used as the measure of performance. There were 14 trials at each speed. In order to maximize difficulty, duration of trial length was increased as the speed of the rotor was increased. At 15 RPM, each trial lasted 20 seconds; at 30 RPM each trial lasted 25 seconds; at 45 RPM each trial lasted 35 seconds; at 60 RPM each trial lasted 45 seconds. Rest periods between trials were inversely related to speed of the rotor. At 15 RPM Ss were allowed a 30 second rest period between trials; a 25 second rest period at 30 RPM, a 20 second rest period at 45 RPM; and a 15 second rest period at 60 RPM. Average number of errors per/second was 2.6 at 15 RPM; 6.9 at 30 RPM; 19.1 at 45 RPM; and 83.4 at 60 RPM.

Ss were instructed to keep the point of their stylus upon the target area as long as the turntable was moving. Each S was given four practice trials at 15 RPM, prior to beginning the task. Because of equipment failure, scores for 14 Ss were lost.

Measures of Expectations and Attraction

A. Probability of Performing Well. (Probability) To measure subjective probability concerning performance, Ss evaluated the following concept using six bipolar adjectives.

Each pair of adjectives were evaluated on a seven point scale; the weight assigned to each point on the scale is shown in parentheses

MY CHANCES OF GETTING A HIGH SCORE

high :	(7)	:	(6)	:	(5)	:	(4)	:	(3)	:	(2)	:	(1)	:	low
weak :	(1)	:	(2)	:	(3)	:	(4)	:	(5)	:	(6)	:	(7)	:	strong
slim :	(1)	:	(2)	:	(3)	:	(4)	:	(5)	:	(6)	:	(7)	:	fat
likely :	(7)	:	(6)	:	(5)	:	(4)	:	(3)	:	(2)	:	(1)	:	unlikely
dark :	(1)	:	(2)	:	(3)	:	(4)	:	(5)	:	(6)	:	(7)	:	bright
good :	(7)	:	(6)	:	(5)	:	(4)	:	(3)	:	(2)	:	(1)	:	bad

above. A Probability score for each *S* was obtained by summing weights over all six pairs of adjectives, with a possible range of 6 to 42.

B. *Attraction toward Good Performance.* (Task Attract) As a measure of motivation to perform well, *Ss* also evaluated the following concept, using six bipolar adjectives.

WHAT A HIGH SCORE MEANS TO ME

valuable :	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	worthless
unimportant :	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	important
useful :	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	useless
boring :	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	pleasant
stupid :	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	clever
interesting :	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	_____	:	uninteresting

A scoring procedure similar to the one used for Probability scores was used here.

The above two scales were completed by *Ss* prior to beginning the Anagrams task and Pursuit Rotor task, and following each change in task difficulty. Thus the scales were completed four times for the Anagrams task and five times for the Pursuit Rotor task.

Personality Measures. *Ss* were tested with the Hand Skills Test and Insolence Scale upon completion of the experimental tasks. Intelligence test scores were obtained from each man's Navy records.

1. *Hand Skills Test.* (Persistence) The Hand Skills Test sought to measure motivation to persist beyond minimum standards on tiring tasks. It consists of sequentially numbered boxes in which examinees pencil five tally marks. The test rapidly promotes hand and arm fatigue and is presented to *Ss* as a measure of hand and finger dexterity. It has a one minute practice session and three parts of four minutes each. A "passing score" is announced prior to each of the four minute parts.⁴ Pretesting had established that this score

⁴Instructions for the first four minute part were: "You have four minutes to complete Part Two. Fill as many boxes as you can. Remember each box must be completed before the next one is started. The more you do the better

could be reached by all examinees in the time allowed.

The test seeks to discriminate between those who stop or slow down after the passing score is reached and those who continue to strive. The score used is number completed in part three minus number completed in the practice session. On the basis of this score, Ss were divided into three equal groups.

2. *Insolence Scale*. The development of this scale has been described in a previous article (Kipnis and Glickman, 1962). It consists of 27 self-description items, which if answered affirmatively would convey the impression of a physically active, aggressive, somewhat hostile and reckless personality, who early in life became independent of family and grade school control, and who it is suspected continues to maintain an independent and rebellious attitude toward most attempts at controlling his behavior. In terms of character structure, the items appear most descriptive of individuals who would be diagnosed as psychopathic personalities.

Items and scoring key for the Insolence Scale may be obtained from the authors. The test is keyed so that high scores denote men high in insolence. Ss were divided at the median of Insolence test scores (10.0) into High and Low Insolent Groups.

3. *General Classification Test (GCT)*. The GCT was used as the measure of general intelligence. This test measures verbal ability and is administered to enlisted men when they first enter the Navy. Ss in this sample were drawn from the middle range of GCT scores, roughly equivalent to IQs from 95 to 115. Using a median split ($GCT = 53$) Ss were divided into High and Low GCT groupings.

Intercorrelation of Tests. The product-moment intercorrelations of the three tests, for the High GCT and Low GCT groupings separately, and for the total sample were computed. The three tests were relatively independent of each other at both levels of intelligence and in the total sample. The range of intercorrelations was from $-.02$ to $+.17$, with a median intercorrelation of $.02$.

Task Order. A counter-balanced design was employed so that about half the Ss taking the anagrams, and half the Ss taking the

your score. In order to pass this part of the test you must fill 100 boxes. Less than 100 is a failing score." Subsequent parts of the test raised passing scores by five, and limited instructions to only announcing a new passing score and the time allowed, e.g. "For this part of the test you must make at least 105 to pass. You have four minutes to finish Part Three."

pursuit rotor task, began work at the hardest level of task difficulty and worked down to the easiest level. For the anagrams task this meant starting with the difficult anagrams and ending with the easy anagrams. For the pursuit rotor task, this meant starting at 60 RPM first and ending with 15 RPM. It should be noted that all Ss expected to work on easy tasks first, on the basis of the practice problems they completed.

Since previous studies had suggested no interaction effects between Persistence and the Insolence Scale, and the present study found no correlation between the two tests, analysis of the results for each test was done separately.

Analysis and Initial Results

To test the hypothesis that task difficulty was the basis for the interaction between the tests and task performance, the results were initially analyzed using a repeated measurement analysis of variance design (case II) described by Winer (1962), with adjustment made for unequal cell entries. The findings provided no support for the hypotheses concerning task difficulty. Neither Persistence nor the Insolence Scale significantly interacted or showed any trends toward interaction with task difficulty. Further, the higher order interaction between task difficulty, intelligence and the tests were also not significant. The results however did yield interactions between each of the two tests and intelligence. To simplify analysis therefore, a single score was obtained for each S on each task and each task rating, by summing his scores over all levels of task difficulty. These total scores were then used in all subsequent analyses.

The initial analysis of the results also indicated that task order had significant effects upon performance. It appears that starting with a hard task and gradually working down to an easier task was disruptive of Ss performance. Ss working in this task sequence solved significantly fewer anagrams ($p < .05$) and made more errors on the pursuit rotor ($p < .01$) than did Ss who started at easy levels of performance and gradually worked up to the harder levels. Therefore, where the results yielded significant interactions between either of the two personality measures and task order, the findings will be presented separately by task order. Otherwise, results for both orders will be combined.

Results for Persistence

a. *Anagrams.* There was a significant interaction ($F = 4.07$, df 2 and 128, $p < .05$) between Persistence, GCT, and total number of anagrams solved. Table 1 presents the average number of anagrams solved by GCT level and Persistence. It can be seen that Persistence was positively related to performance among Ss with low ability and negatively related among Ss with high ability.

TABLE 1
Average Number of Anagrams Solved: Ss Grouped by GCT and Persistence
($N = 140$)

Persistence	High GCT		Low GCT		<i>t</i>	<i>P</i>
	<i>N</i>	\bar{X}	<i>N</i>	\bar{X}		
High	(21)	80.64	(25)	84.89	1.00	ns
Middle	(26)	101.50	(22)	68.19	3.77	.01
Low	(30)	95.14	(16)	76.50	1.97	.05

Comparison of individual means (see Table 1) suggested that the basis for the significant interaction was that intelligence facilitated performance among Ss classified as middle or low in Persistence, but not among Ss classified as high in Persistence.

b. *Pursuit Rotor.* The main result was that among Ss who started at 60 RPM and worked down to 15 RPM, fewer errors were made by those classified as high in Persistence than were made by those classified as middle or low ($p < .01$). Persistence was not related to pursuit rotor performance among Ss who started at 15 RPM and ended at 60 RPM.

Persistence and Task Attraction

While the finding for the anagram task provides support for the contention that persistence interacts with intelligence, it does not provide any cues as to the basis for this effect. The present results seem to rule out the possibility that task difficulty per se is the basis for this interaction. An alternate possibility is that the more intelligent person is not as interested in the work assigned to him as is the less intelligent person assigned similar work, and hence is not motivated to try very hard. The hypothesis here then would be that Persistence facilitates performance among individuals who are attracted to a given goal.

As a preliminary test of this possibility, Ss were classified in terms of their Task-Attract scores made prior to beginning the

anagram task (using a median split for Pre-Task Attract scores). Separate product-moment correlations were then computed between Persistence scores and number of anagrams solved for Ss classified as being high in Pre-Task Attract scores and for Ss classified as being low in Pre-Task Attract scores. This analysis was limited to Ss who started with the easy anagrams first, since the consequences of starting with the hard anagrams had unexpected and not too well understood effects on subsequent performance and task motivation. Among Ss who were above the median on the Pre-Task Attract scores, the product-moment correlation between total number of anagrams solved and Persistence was $+ .33$ ($p < .05$, $df\ 38$). Among Ss below the median on the Pre-Task Attract scores, the product-moment correlation was $- .51$ ($p < .01$, $df\ 33$). Using a z' transformation, the difference between these two correlations was reliable beyond the .01 level. While there were not enough cases to also compute correlations at each GCT level, inspection of scores for high and low GCT groupings parallel the combined results.

A similar analysis was done for pursuit-rotor performance using Ss who started at 15 RPM. For this task, results similar to those found for the anagram task were obtained for the earlier stages of performance at 15 and 30 RPM, although the results were not significant. Among Ss with above the median Pre-Task Attract scores, the correlation between Persistence and number of pursuit-rotor errors was $-.10$ ($P\ ns$, $df = 34$). The similar correlation among Ss with below the median Pre-Task Attract scores was $+.23$ ($P\ ns$, $df = 31$).

Discussion—Persistence

Perhaps the most suggestive results were based upon the analysis of the relationship between task attraction, Persistence, and number of anagrams solved. While these results require experimental verification, they suggest that to the extent an individual is motivated to strive, high Persistence scores may ensure that his efforts will be successful.

The results for individuals with low Persistence scores appear similar to the reported effects of anxiety upon complex task performance (e.g. Taylor and Spence, 1952) or the effects of stress upon persons with high task drive (Longenecker, 1962). What is called to mind in the case of those with low Persistence scores are

persons who perhaps do their best when they are not personally involved in the outcome, but cannot mobilize their efforts efficiently when attempting to achieve a desired goal. Studies by Katz and his associates (1963) on the debilitating effects of heightened motivations on negro's test performance also seems related to the effects found here.

Results for the Insolence Scale

a. *Anagrams.* The only significant relationship between the Insolence Scale and the Anagrams task was on Ss ratings of their attraction toward the task. It may be recalled that the Task-Attract score was obtained by summing the four ratings of task attraction into one score. Analysis of this summed Task Attract scores yielded a significant second-order interaction between Insolence, GCT, and Task Order ($F = 4.90$, df 1 and 132, $p < .05$). Since at best, second-order interactions are difficult to interpret, the data were re-analyzed separately by task order. This analysis indicated that among Ss who started with easy anagrams first, there was a significant interaction between Insolence, GCT, and Task Attract scores ($F = 8.40$, df 1 and 67, $p < .01$). As shown in Table 2, among Ss above the median in intelligence, those high in Insolence had lower Task Attract scores than did subjects classified as low in Insolence. The reverse of this relationship held for Ss who were below the median in intelligence. This relationship is similar to those found in the prior field studies in which the dependent variables were superior's evaluations of performance.

Among Ss who started with the hard anagrams first, (see Table 2) results of the analysis of variance were that Ss high in insolence had significantly lower Task Attract scores than did Ss low in in-

TABLE 2
*Motivation to Do Well on Anagram Task (Task Attract): Ss
Grouped by Insolence Scale, GCT, and Task Order*

	Easy Anagrams First ($N = 71$)				Hard Anagrams First ($N = 69$)			
	High GCT		Low GCT		High GCT		Low GCT	
	N	\bar{X}	N	\bar{X}	N	\bar{X}	N	\bar{X}
High Insolence	(18)	135.80	(15)	152.76	(18)	134.38	(14)	121.20
Low Insolence	(21)	141.96	(17)	125.82	(20)	148.81	(17)	137.41

solence ($F = 7.96$, df 1 and 65, $p < .01$). This was true at both levels of intelligence.

b. *Pursuit Rotor*. The relationship between Insolence, GCT, and Task Attract scores followed the same pattern found in the anagram task. Among Ss starting at 15 RPM, there was a significant interaction between GCT and Insolence. That is, Insolence Scale scores were negatively related to Task Attract scores among the more intelligent and positively related to Task Attract scores among the less intelligent ($F = 5.18$, df 1 and 61, $p < .05$). There was no statistically significant relationship between Insolence and Task Attract scores among Ss who started the pursuit rotor task at 60 RPM.

One final set of additional data, of an accidental nature, was obtained from the pursuit-rotor task which related to Insolence Scale scores. To facilitate data collection, four Ss were run simultaneously in the same room, each separated from the others by cloth hospital screens. It was possible to observe each of the Ss through a one way mirror; although Ss could not see each other without pushing aside the screens. Instructions given at the beginning of the task specifically requested that no talking be done at any time. These instructions were usually repeated during the task as well. During the first week of data gathering it became obvious that many of the groups held conversations, despite repeated requests by the experimenters not to talk. Since the experimenters could not "beat them," they decided to gather data on the frequency with which each S talked during the experiment. A simple tally was kept on whether or not each man talked during each trial and during each rest period. The same prohibitions against talking were continued.

Observations were obtained on a total of 64 Ss; 34 who started at 15 RPM and 30 who started at 60 RPM. Among those starting at 15 RPM there was a significant interaction between Insolence and GCT, with reference to frequency of talking ($F = 4.83$, df 1 and 30, $p < .05$). Among the more intelligent, high insolent Ss spoke more frequently than did low insolent Ss; the reverse of this relationship was true among the less intelligent. While there was no statistically significant relationship between insolence scores and talking among Ss starting at 60 RPM, the general trend was for high insolent Ss to talk more than low insolent at both levels of intelligence. ($F = 2.44$, df 1 and 26, p ns).

Discussion—Insolence Scale

The behavior of Ss classified as both high in insolence and high in intelligence is consistent with the general behaviors one might expect from persons diagnosed as psychopathic personalities. Several investigators (Rasmussen, 1961; Whitman, Trosman, and Koenig, 1954) have suggested that such personalities encounter their main difficulties in their unwillingness, rather than their inability, to accept the social role which is required of them. This unwillingness was indicated in the present study in the expressed lack of interest of High Insolent-High Intelligent Ss in doing well on both the pursuit rotor and anagram tasks, and in their continued conversations during the pursuit rotor task, despite prohibitions against talking by the experimenter. While the Insolence Scale was not related to task performance over the short periods of time used here, it does seem probable that over longer periods of time, these negative attitudes toward striving would manifest themselves in poorer performance, especially in situations involving interactions with authority figures.

The results then parallel in a lab setting those of the original field studies. Unfortunately, the lack of support of the task difficulty hypothesis provides no additional understanding of the basis of these interactions. It is not possible to accept the alternate hypothesis of a fixed relationship between insolence, intelligence, and behavior. This is because when Ss were required to begin on the hard anagram task first, all high insolent Ss had low Task Attract scores, regardless of their intelligence level.

In any case, it appears that under certain conditions, not isolated by this study, persons with high Insolence Scale scores will be motivated to strive for socially acceptable goals. In general these conditions seem most effective with individuals who are average to below-average in intelligence. However, if we assume that persons with high Insolence Scale scores have similar character structures regardless of their intelligence levels, then it follows that there should be incentive conditions which might also induce High Insolent-High Intelligent individuals to strive for socially acceptable goals.

Summary

Prior field studies had found that intelligence moderated the validity of two non-cognitive tests that had been developed to predict the job performance of Navy enlisted men. The first test was an attempt to measure persistence, and the second test was a purported measure of passive-aggressive character structure called the Insolence Scale. The present study tested the hypothesis that task difficulty was the basis for the field study results. From 70 to 140 Navy enlisted men completed three tasks, each of which was varied in task difficulty. Results did not support the hypothesis concerning task difficulty. However, they did yield significant interactions between intelligence, the two tests, and aspects of task performance. The findings then parallel in a lab setting those of the original field studies. Internal analysis suggested that task motivation, rather than task difficulty, may be the basis for the interactions between the persistence test and performance.

REFERENCES

- Atkinson, J. W. "Motivational Determinants of Risk-Taking Behavior." *Psychological Review*, LXIV (1957), 359-372.
- Feather, N. T. "The Relationship of Expectation of Success to Reported Probability Task Structure, and Achievement Related Motivation." *Journal of Abnormal and Social Psychology*, LXVI (1963), 231-238.
- Fiedler, F. E. "Leaders Attitude, Group Climate and Group Creativity." *Journal of Abnormal and Social Psychology*, LXV (1962), 308-318.
- Haythorn, W., Couch, I., Haefner, P., Langham, G., and Carter, L. F. (1956) "The Effects of Varying Combinations of Authoritarian and Equalitarian Leaders and Followers." *Journal of Abnormal and Social Psychology*, LIII (1956), 210-219.
- Katz, I., Epps, E. G., and Axelson, L. "Effect upon Negro Digit-Symbol Performance of Anticipated Comparison with Whites and with Other Negroes." ONR Technical Report #5, Contract 285 (24), 1963.
- Kipnis, D. "A Non-cognitive Correlate of Performance among Lower Aptitude Men." *Journal of Applied Psychology*, XLVI (1962), 76-80.
- Kipnis, D. "The Relationships between Persistence, Insolence, and Performance, as a Function of General Ability." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXV (1965), 95-110.
- Kipnis, D. and Glickman, A. S. "The Prediction of Job Performance." *Journal of Applied Psychology*, XLVI (1962), 50-56.
- Longenecker, E. D. "Perceptual Recognition as a Function of An-

- xiety, Motivation, and the Testing Situation." *Journal of Abnormal and Social Psychology*, LXIV (1962), 215-221.
- McGrath, J. E. "The Influence of Positive Interpersonal Relations on Adjustment and Effectiveness in Rifle Teams." *Journal of Abnormal and Social Psychology*, LXV (1962), 365-375.
- Mandler, G. and Sarason, S. B. "A Study of Anxiety and Learning." *Journal of Abnormal and Social Psychology*, XLVII (1952), 166-173.
- Mayzner, M. S. and Tressalt, M. E. "Anagram Solution Timed: A Function of Letter Order and Word Frequency." *Journal Experimental Psychology* LVI (1958), 376-379.
- Meyers, A. "Team Competition, Success, and the Adjustment of Group Members." *Journal of Abnormal and Social Psychology*, LXV (1962), 325-332.
- Osler, S. F. and Trautman, G. E. "Concept Attainment: II. Effect of Stimulus Complexity upon Concept Attainment of Two Levels of Intelligence." *Journal of Experimental Psychology*, LXII (1961), 9-13.
- Rasmussen, J. E. "An Experimental Approach to the Concept of Ego Identity as related to Character Disorder." Unpublished doctoral dissertation thesis. American University, Washington, D. C., 1961.
- Saunders, D. R. "Moderator Variables in Prediction." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVI (1956), 209-222.
- Taylor, J. A. and Spence, K. W. "The Relationship of Anxiety Level to Performance in Serial Learning." *Journal of Experimental Psychology*, XLIV (1952), 61-64.
- Whitman, R. M., Trosman, H., and Koenig, R. "Clinical Assessment of Passive-Aggressive Personality." *A.M.A. Archives of*
- Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.

THE SOCIAL DESIRABILITY VARIABLE: IMPLICATIONS FOR TEST RELIABILITY AND VALIDITY

ALFRED B. HEILBRUN, JR.¹

Emory University

A full decade has passed since Edwards directed attention to the social desirability (SD) variable in personality testing by publishing his own SD-"controlled" test (1954), and the interested bystander has a virtual mountain of empirical data from which he may presumably draw conclusions. Most of the studies which have been published bearing upon SD have sought to discover whether normals tend to describe themselves in socially desirable ways on personality tests (which they do) or whether normals can manipulate their responses to test items in socially desirable or undesirable ways if instructed to do so (which they can).

One noteworthy trend suggested by the SD evidence is that when the investigation has been limited to studying the relationships of some SD measure to personality test scores or to measuring changes in scores under instructional sets, SD typically has been interpreted as a response set and an undesired source of performance variance. Factor analytic investigations of personality tests tend toward the same conclusion, although some factor analysts (e.g., Jackson and Messick, 1962) have adopted the intermediate position that SD is a "style" of performance which may affect both test and nontest behavior. On the other hand, in the minority of studies which have engaged in experimental manipulation of nontest variables (e.g., Marlowe, 1962; Marlowe and Crowne, 1961) or have employed some nontest criteria (e.g., Heilbrun,

¹ This study was conducted while the investigator was a Visiting Associate Professor, Department of Psychology, University of California, Berkeley, Calif.

1962; Heilbrun and Goodstein, 1961), serious doubts have been raised regarding the interpretation of SD as a response set and, ipso facto, a source of error on personality tests.

The interpretive problem can be traced to what seems to have been, in retrospect, a logical flaw of one-decade vintage. Edwards (1953) assumed that a high positive correlation between the SD ratings for specified behaviors and the tendency to endorse these behaviors as self-characteristic indicated a massive response set to answer items in socially desirable ways among college students. Subsequent studies which have been confined to purely psychometric relationships have usually accepted this logic without serious questioning. The fact that Marlowe and Crowne and their associates have reported a series of experimental studies (Crowne and Liverant, 1963; Crowne and Strickland, 1961; Marlowe, 1962; Marlowe and Crowne, 1961; Salmon and Crowne, 1962) in which their own measure of SD predicted a variety of socially conforming behaviors opens the way to the opposite logical proposition. *Persons who are more sensitive to how their behavior will be accepted by others (high SD Ss) are more likely to conform to standardly employed test instructions to answer frankly and honestly.* It would be anticipated that high SD would lead to a more factual self-description based upon this proposition rather than to the positive dissembling suggested by the SD response set assumption.

There are several studies which lend support to the contention that the tendency to respond in socially desirable ways is positively related to test validity. Heilbrun and Goodstein (1961) found that when the Edwards Personal Preference Schedule (EPPS) was specifically "tailored" for each S so that the individual had even less opportunity to select item alternatives based upon differential SD than offered by the standard form, predictive validity was reduced. Heilbrun (1962) also demonstrated that the Need Achievement scale of the EPPS (involving some curtailment of the effect of individual differences in SD responding) was predictively inferior to a similar scale with no SD controls. Finally, Smith (1959) has shown that the MMPI K scale, derived as an indicant of positive dissembling among psychiatric patients, was positively related to a measure of self-insight for normal groups. The greater S's tendency to endorse socially desirable behavior, the greater self-insight attributed to him by others.

Thus, the first hypothesis tested in the present study was: *I. A personality scale will be more valid for high SD Ss than for low SD Ss.*

The second purpose of the present study was to consider the relationship between SD responding and personality test reliability (i.e., test-retest stability). This interest was aroused by an inadvertent finding in a prior study (Heilbrun, 1964) that a positive relation held between a measure of SD and a measure of "role consistency." The latter metric, devised by Block (1961), indicates the similarity of self-description over several difference interpersonal situations and may be taken as an estimate of stability of the person's self-concept. *S* was asked, for example, what he is like when interacting with a boss, his mother, a girl friend, etc.? One interpretation of this positive relationship would be that *Ss* who are more cautious in self-description (high SD) tend to embrace the same positive attributes as self-descriptive for each interpersonal situation and to deny negative attributes thereby providing a falsely stable self-conceptual picture. This would be an SD response set interpretation. On the other hand, it could be argued that the person who has a more positive self-concept is more likely to have a more stable self-concept. This would follow from the assumptions that socially reinforced behaviors will show a greater habit strength and that the greater strength of social behaviors will, in turn, provide the foundation for a more stable self-concept. Empirical support for the latter interpretation is provided by Gough (Gough and Heilbrun, 1965). One hundred men were administered a checklist of adjectives under self-description instructions on two occasions some six months apart as part of an assessment program. These men were also rated by 10 staff observers on the same checklist. Comparison of the ratings for those men whose self-descriptions were most stable with those who were less stable in self-description indicated clear differences in positive and negative attributes. The "reliable" man was rated as more quick, cheerful, confident, cooperative, energetic, enterprising, fair-minded, foresighted, friendly, informal, ingenious, insightful, progressive, spontaneous, and versatile. In contrast, the "unreliable" man was observed to be more awkward, painstaking, arrogant, mild, prejudiced, resentful, smug, and unassuming.

Since test-retest stability of objective personality measures re-

flects the degree of relationship between one self-concept obtained from a given *S* at time 1 and another self-concept obtained from him at a later time 2, a reliability coefficient can be considered an index of self-conceptual stability. (Other factors also contribute to a less-than-perfect correlation, of course.) Based upon the proposition that higher SD *Ss* will have more stable self-concepts than low SD *Ss*, the second hypothesis tested in this study was: II. *High SD Ss will show less change in their self-descriptions following the receipt of new self-relevant but biased information than low SD Ss.*

Method

Subjects. Forty-four male volunteers from large undergraduate classes at the University of California, Berkeley, were used in the experimental part of this study. Fifty-two more *Ss*, similarly obtained, including 29 males and 23 females, were added later.

Measures. The Adjective Check List (ACL) (Gough and Heilbrun, 1965) was employed as the personality measure. The ACL includes 300 common behavioral adjectives, and the standard instructions request *S* to check those which he considers self-descriptive.

The Defensiveness (Df) Scale. Heilbrun has reported the derivation elsewhere (1961) of a scale scored from the ACL which measures the tendency to portray oneself in a favorable way on the ACL. It was derived empirically by determining the adjectives which differentiated maladjusted males who described themselves favorably from those maladjusted males whose self-descriptions were more in accord with their level of adjustment and which held up upon replication. These adjectives have been T-scaled so that the college mean is set at 50 with $SD = 10$; higher scores indicate greater test-taking defensiveness. The Df Scale correlates .64 with the K scale of the MMPI for maladjusted college males and .35 for normal college males.

Like all other SD-type scales, Df confounds *S*'s approach to test responding and his actual level of adjustment. High scores, for example, could indicate a "fake-good" attitude, a superior level of adjustment or both. In any case, high Df scorers are clearly those males who have endorsed behaviors which are socially desirable whereas low Df scorers have portrayed themselves in a less socially desirable way on the ACL.

The Abasement (Aba) Scale. The derivation of an ACL scale to measure abasing tendencies has been described in another study along with experimental evidence supporting scale validity (Heilbrun, 1959). The definition (Edwards, 1954) of abasement provided judges to govern their selection of ACL adjectives for the scale was as follows.

• "To feel guilty when one does something wrong, to accept blame when things do not go right, to feel that personal pain and misery suffered does more good than harm, to feel the need for punishment for wrong doing, to feel better when giving in and avoiding a fight than when having one's own way, to feel the need for confession of errors, to feel depressed by inability to handle situations, to feel timid in the presence of superiors, to feel inferior to others in most respects."

Role Consistency (RC). This measure of self-conceptual stability was included in the study, although it was not crucial to either hypothesis under test. Its inclusion did allow for an attempted replication of the previously obtained positive relationship between Df and RC.

The RC measure requires *S* to rank order a standard set of 20 adjectives from most to least descriptive for each of eight interpersonal situations (e.g., "with someone in whom you are sexually interested," "with your employer or someone of equal status," "with your father," "with your mother," etc.). The RC score is given by the Kendall coefficient of concordance (Siegel, 1956, pp. 229-238), a type of multiple rank-order correlation which ranges from .00 to 1.00. A high score indicates that *S* is assigning similar ranks to each adjective (i.e., showing the same pattern of behavior) for all of the interpersonal roles.

Information. The information provided to the *Ss* to elicit self-descriptive change was an edited verbatim typescript of a speech² by an Army psychiatrist, Major William E. Mayer. In the original speech Major Mayer made many derogatory comments about the attributes of 18-22 year-old American males based upon their behavior during the Korean police action. Special emphasis was

² The writer would like to express his appreciation to Dr. Gordon Polder who made the edited version available for this study. Specific time and place of the lecture is unknown, but the reader is referred to the February 24, 1956 issue of *U. S. News & World Report* for essentially the same information.

placed upon the glaring failures of the American prisoners-of-war as indicated by their complete capitulation to the captor and the loss of such basic values as loyalty and discipline. In general the young American male was pictured as passive, effeminate, selfish, dishonorable, dependent, and lacking initiative; the 11-page typescript was a completely negative one following the editing.

Procedure. The experimental Ss were seen individually by *E*. Each was first given the RC and ACL measures to complete (in that order) with no explanation of purpose. Upon completion of RC and ACL-1, *S* was provided a copy of the edited typescript with instructions to "read it quickly but try to digest what he has to say."

When *S* had finished reading the typescript he was given instructions by *E* intended to facilitate self-descriptive change. (This was considered necessary to avoid the possibility of insufficient change in adjective endorsement patterns after the 10-15 minute test-retest interval to allow the hypotheses to be tested.) The essential features of these instructions were: (1) "we knew" that in adjective checking there are many marginal adjectives (i.e., checked as characteristic but barely so, not checked but barely non-characteristic); (2) *S* had just been given some opinions by an expert regarding young men of his age; and (3) *S* was to fill out the ACL again "keeping in mind what Major Mayer had said to whatever extent you feel it might apply to you."

The entire experimental procedure took about one hour for most Ss. Filling out of the measures and reading of the typescript were performed in a room separate from *E*'s location during the experiment.

Results

Experimental Effect. It was of interest to determine whether there was any systematic effect of the experimental condition upon adjective self-description for the entire group of 44 Ss prior to testing the specific hypotheses. Since there was no control group, however, none of these general findings can be clearly attributed to the effects of the intervening condition.

There was a mean of 111.1 adjectives checked on ACL-1 as self-characteristic, whereas the group endorsed an average of 106.4 adjectives on ACL-2, a rather inconsequential drop considering the

large variability for each count ($SD = 31.8$ and 34.2 , respectively). There was a mean of 41.1 changes in endorsement ($SD = 14.8$) which includes both drops (i.e., adjectives checked on ACL-1 but not on ACL-2) and additions (i.e., adjectives checked on ACL-2 but not on ACL-1).

Qualitative evaluation considered whether the changes in adjective endorsement were of a generally favorable or unfavorable character. Gough (Gough and Heilbrun, 1965) has provided the 75 ACL adjectives judged to reflect most favorably on the endorser and the 75 judged to reflect most unfavorably. A favorability ratio (FR) was derived which would portray the tendency toward self-enhancement or self-depreciation as follows:

$$FR = \frac{(\text{No. Fav 2} - \text{No. Fav 1}) - (\text{No. Unfav 2} - \text{No. Unfav 1})}{\text{Total Number Changes in Endorsement}}$$

Thus, an increase in the number of endorsed adjectives keyed as favorable or a decrease in endorsement of adjectives keyed as unfavorable on ACL-2 (as compared to ACL-1) contributes positively to the numerator, the opposite negatively. Dividing by the total change score removes the effect of differential susceptibility to self-descriptive change. The potential range of scores is from -1.00 to $+1.00$. The overall FR mean was $.09$ ($SD = .16$) which indicates that following the experimental condition self-conceptual changes tended to be in a positive direction with nine percent more of the changed endorsements being in a favorable direction than in an unfavorable direction. FR scores ranged from $-.28$ to $+.46$.

Hypothesis I. A personality scale will be more valid for high SD Ss than for low SD Ss. Since high abasing Ss should be more prone to incorporate the unfavorable observations presented in the experimental condition than low abasing Ss, it was predicted that the Abasement scale should relate negatively to FR. Using the college mean of $T = 50$ ($SD = 10$) as the cutting point, 26 high abasers and 18 low abasers were defined. The Abasement scale was nonpredictive of the FR criterion, since the FR mean score for high abasers ($.08$) was almost identical to that provided by low abasers ($.09$ ($t = .21$ for 42 df ; $p > .80$)).

The deviation of FR scores from $.00$ was employed as the dependent variable in testing Hypothesis I. High abasers were assigned plus accuracy scores if FR was below $.00$ (was self-derogatory) and

minus scores if FR was above .00 (was self-enhancing); the opposite pattern of scoring plus and minus accuracy scores held for low abasers. The mean Abasement scale accuracy score for the 18 high ($T > 50$) Df Ss was .07 ($SD = .21$) and for the 26 low Df Ss the mean was $-.07$ ($SD = .19$). This difference was significant ($t = 2.27, p < .05$ for 42 df) and supported Hypothesis I.

Hypothesis II. High SD Ss will show less change in their self-descriptions following the receipt of new self-relevant but biased information than low SD Ss. Product-moment correlations between Df and RC and the criterion of post-experimental self-descriptive change (total adjectives changed) are reported in Table 1. It can

TABLE 1
*Correlations between Df and RC and Three Criteria of
Post-Experimental Self-Descriptive Change*

	RC	Total No. Adjectives Changed
Df	.42**	-.34*
RC	—	-.31*

* $p < .05$.

** $p < .01$.

be noted that the previously reported positive relationship between Df and RC was replicated in the present findings. Support for Hypothesis II was found in the significant negative correlation between Df and the total change score. The higher the Df score, the more stable the self-concept in the face of unfavorable information relevant to S's age group.

Since it might be reasonably contended that the experimental method by which test-retest stability was measured in this investigation represents a radical departure from the usual conditions under which test reliability is estimated, an additional empirical study was conducted. A sample of 52 college undergraduates (29 males and 23 females) were group-administered the ACL under standard instructions and 10 weeks later were again tested with the ACL under standard conditions. In line with Gough's procedure, these 52 Ss were separated into "reliable" responders and "unreliable" responders on the basis of their total adjective change score. Male and female distributions were cut separately as near as possible to their midpoints so that an approximately equal number of

males and females would be included in the "reliable" and "unreliable" groups. The former group included 15 males and 10 females and the latter group was made up of 14 males and 13 females. Cutting scores for inclusion as a "reliable" responder was < 56 adjectives changed for males and < 44 adjectives changed for females. Based upon the experimental stability findings, it would be predicted that the "reliable" responders would be higher on Df than the "unreliable" responders. The Df comparison (Table 2)

TABLE 2

Comparison of Df Scores for "Reliable" and "Unreliable" ACL Test Responders

"Reliable" Test Responders			"Unreliable" Test Responders			<i>t</i>
<i>N</i>	Mean Df	SD	<i>N</i>	Mean Df	SD	
25	53.9	9.0	27	46.2	7.1	3.34***

*** Significant at the .001 level of confidence.

clearly supported this prediction and allows generalization of the experimental stability results to standard reliability measurement.

Discussion

The results of the present study represent one more illustration of the fact that when the functional boundaries of the social desirability variable are scrutinized by methods other than within-test correlation or instructional directive to fake, a far different interpretation of SD emerges than that of a response set and a source of predictive error. The present investigation was addressed to the possibility that Edwards' initial assumption that SD should function as a response set in personality assessment and, accordingly, as a source of predictive error was incorrect. The experimental findings do point towards the opposite conclusion. Young college males whose self-descriptions are of a more socially desirable character provided test records which were both more valid and reliable. One possible explanation of these findings, offered earlier, would be that high SD persons are more likely to conform to the instructional sets under which personality questionnaires are administered which typically urge the respondent to answer items as accurately or honestly as possible. Smith's conclusion that higher SD normals are more self-insightful represents a second way of accounting for the present validity findings.

One further comment regarding the validity findings is required. The present results were obtained using a rationally-derived personality scale. Such scales typically are more obvious in contentual meaning to *S* than those constructed by empirical techniques and, accordingly, are more vulnerable to the degree of cooperation and level of self-insight of the individual tested. Further research is necessary to determine whether the present experimental validity findings would hold for other than rational scales.

A second major finding of this study was that high SD *Ss* demonstrated less change in self-description than low SD *Ss* when experimentally confronted with unfavorable information relevant to their age group. Findings for an independent sample of *Ss* indicated that the greater stability of the high SD test responder can clearly be generalized to reliability measurement as standardly obtained. The positive and replicated relationship between *Df* and a measure of self-conceptual stability offers a reasonable basis for expecting this result. The self-concept would be expected to vary in kind with the type of social behaviors performed and in stability with the strength of social habits. For example, a person who dominates others in a high proportion of interpersonal situations is more likely to conceive of himself as a dominant person at two points in time than a person who less frequently adopts a dominant role. Furthermore, it would be expected that the latter person would be more susceptible to self-conceptual change if an experience intervening between time 1 and time 2 provided evidence counter to his initial self-concept. This source of test unreliability would be especially likely among adolescents and young adults who are experiencing the transition from childhood to adult behavioral roles.

If further research supports the conclusion of this study, namely, that the tendency to perform on tests (and presumably in nontest situations) in a socially desirable way is a requisite for test validity and reliability rather than a source of error, it will point a way toward reversing the trend of increasing skepticism toward and even legal legislation against personality tests. Instead of proliferation of tests bearing the same intrinsic flaws, greater ingenuity must be directed toward clarifying the limitations upon predictive potential of all tests. The present results suggest that one step in this direction would be to acknowledge the possibility that even within a grossly normal population personality questionnaires may not be useful for

a substantial proportion of individuals among which are those whose behaviors tend to be socially undesirable.

Summary

The logical assumption which has influenced social desirability (SD) research for a decade is that SD acts as a response set in personality testing and as a source of predictive error. The present study proposed the opposite, namely, that high SD responders should provide more valid and reliable test records. The Adjective Check List (ACL) was administered to 44 college males followed by negative information relevant to S's age group. Upon readministration of the ACL, it was found that the prediction of a more negative ACL-2 self-description for high abasing Ss was supported for high SD Ss only; the reverse relationship held for low SD Ss. High SD Ss were also more stable in their self-descriptions than low SD Ss, and this finding was replicated for an independent group tested under standard test-retest reliability conditions.

REFERENCES

- Block, J. "Ego Identity, Role Variability, and Adjustment." *Journal of Consulting Psychology*, XXV (1961), 392-397.
- Crowne, D. P. and Liverant, S. "Conformity under Varying Conditions of Personal Commitment." *Journal of Abnormal and Social Psychology*, LXVI (1963), 547-555.
- Crowne, D. P. and Strickland, Bonnie. "The Conditioning of Verbal Behavior as a Function of the Need for Social Approval." *Journal of Abnormal and Social Psychology*, LXIII (1961), 395-401.
- Edwards, A. L. "The Relationship between Judged Desirability of a Trait and the Probability that the Trait Will Be Endorsed." *Journal of Applied Psychology*, XXXVII (1953), 90-93.
- Edwards, A. L. *Manual for the Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1954.
- Gough, H. G. and Heilbrun, A. B. *Manual for the Adjective Check List and the Need Scales for the ACL*; Palo Alto: Consulting Psychologists Press, 1965.
- Heilbrun, A. B. "Validation of a Need Scaling Technique for the Adjective Check List." *Journal of Consulting Psychology*, XXIII (1959), 347-351.
- Heilbrun, A. B. "The Psychological Significance of the MMPI K Scale in a Normal Population." *Journal of Consulting Psychology*, XXV (1961), 486-491.
- Heilbrun, A. B. "Social Desirability and the Relative Validities of Achievement Scales." *Journal of Consulting Psychology*, XXVI (1962), 382-386.

- Heilbrun, A. B. "Conformity to Masculinity-Femininity Stereotypes and Ego Identity in Adolescents." *Psychological Reports*, XIV (1964), 351-357.
- Heilbrun, A. B. and Goodstein, L. D. "Social Desirability Response Set: Error or Predictor Variable?" *Journal of Psychology*, LI (1961), 321-329.
- Jackson, D. N. and Messick, S. "Response Styles on the MMPI: Comparison of Clinical and Normal Samples." *Journal of Abnormal and Social Psychology*, LXV (1962), 285-299.
- Marlowe, D. "Need for Social Approval and the Operant Conditioning of Meaningful Verbal Behavior." *Journal of Consulting Psychology*, XXVI (1962), 79-83.
- Marlowe, D. and Crowne, D. P. "Social Desirability and Response to Perceived Situational Demands." *Journal of Consulting Psychology*, XXV (1961), 109-115.
- Salmon, A. R. and Crowne, D. P. "The Need for Approval, Improvisation, and Attitude Change." Paper read at Midwestern Psychological Association, Chicago, May, 1962.
- Siegel, S. *Nonparametric Statistics*. New York: McGraw-Hill, 1956.
- Smith, E. E. "Defensiveness, Insight, and the K Scale." *Journal of Consulting Psychology*, XXIII (1959), 275-277.

SITUATIONAL AND INDIVIDUAL DETERMINANTS OF ATTITUDE SCALE RESPONSES

MARION STEININGER¹
Moravian College

Two related issues recurrently face psychology. One is how to analyze all behavior as a function of both organismic and environmental characteristics. Several approaches have been explored in detail, notably in Lewin's theoretical articles (Lewin, 1935), Murphy's bio-social personality theory (Murphy, 1947), Brunswik's monograph outlining his "probabilistic functionalism" (Brunswik, 1947), and Orne's social-psychological interpretation of experimental situations (Orne, 1962).

In spite of such efforts, the basic notion that $B = f(P, E)$ appears to be ignored in many interpretations of questionnaire or self-report inventory data. If Lewin's formulation were consistently applied, then responses to questionnaires would be understood to reflect both characteristics of the person, like "traits" and "attitudes" (P), and the testing situation, including the specific questions (E). Or, to put it differently, theorizing about personal characteristics (P) would consistently take into account both sets of observables: the responses (B) and the testing situation (E).

Typically, however, "traits" and "attitudes" are inferred from questionnaire responses when a standard set of questions has been administered to a suitable group under constant conditions. The rationale appears to be the same one which Brunswik criticized in relation to the classical experimental method in psychology (Brunswik, 1947): since the actual questions and the testing conditions are constant from *S* to *S* (even if all *Ss* do not interpret the

¹ Now at Rutgers University.

questions in the same manner), observed individual differences may be attributed to the Ss. Yet this does not follow, since the "traits" or "attitudes" that are found reflect the specific testing situation, including the specific questions, as much as they do the Ss in the sample. The difference is that the amount of "trait" is varied by using a heterogeneous sample, but there is no variation in the testing situation, since it is held constant.

This is one of the considerations that has led to proposals for construct validation (Cronbach and Meehl, 1955; Campbell, 1960), and has led Campbell and Fiske to argue that claims should not be made about an "honesty" test until it can be demonstrated that "honesty" scores and "honesty" behavior in some other situation are more correlated, in terms of Ss' performances, than "honesty" scores and "intelligence" or "anxiety" scores (Campbell and Fiske, 1959).

The second recurrent issue in psychology is to what extent personal characteristics, like traits, are "general," and to what extent they are "specific" to particular situations (Hartshorne and May, 1928; Allport, 1937; Burton, 1963). These two issues may be seen to be intimately related when they are approached field-theoretically. It was suggested above that any score must be understood to reflect both the testee's characteristics and those of the test situation. If this approach is extended to the second issue, then it would seem to follow that whether a trait appears to be general or specific depends on how much variation there is in the sample of Ss, and how much variation there is in the sample of situations or questions.

Specifically, the hypothesis to be tested is that for a theoretically derived subgroup of questions, there will be variation in responses as a function of variations in the Ss, and that for given subgroups of Ss, there will be variations in responses as a function of variations in the questions. In other words, responses to self-report questions will vary systematically as a function of variations in both Ss and questions.

Method

Steininger, Johnson, and Kirts investigated unethical behavior in ordinarily ethical persons, using as an example of such behavior cheating on college examinations (Steininger et al., 1964). To

study attitudes about such cheating, they developed a questionnaire which was based on specific situations related by two hypotheses: the extent to which students say that cheating is justified increases as a function of both increased anxiety and increased hostility. They hypothesized further that these are more strongly induced by "poor" professors, "meager, uninteresting" course content, and "hard" and "senseless" tests than by "good" professors, "new and interesting" content, and "easy" and "meaningful" tests. One other comparison was introduced: professor "leaves" versus "stays" during the test. These five dichotomies were combined into 32 factorial situation descriptions, which are presented in the questionnaire in random order. Two sample descriptions are:

The content of the course is generally meager and uninteresting. As a teacher, the professor is generally good. The tests are easy, but based on senseless detail. The professor stays in the room during the tests, reading and looking around.

The content of the course is generally new and interesting. As a teacher, the professor is generally poor. The tests are hard, but sensible and meaningful. The professor leaves the room during tests, returning at the end to collect the papers.

For each situation, the subject had to answer on a five-point scale how justified cheating would be, how much urge he would have to cheat, how much he would copy, how much he would let others copy from him, and how guilty he would feel. The response category most closely related to the concept "attitude" is "justification." For each situation, the *S* could say that justification (*J*) was "very great" (5), "great" (4), "moderate" (3), "slight" (2), or "non-existent" (1).

The situations were classified according to how many separate elements (tests hard, professor good, etc.) were hypothesized to contribute to more rather than less anxiety and hostility. An anxiety-hostility (A-H) index was thus developed for the situations, values on which ranged from zero (course interesting, professor good, tests easy, tests meaningful) to four (course uninteresting, professor poor, tests hard, tests senseless).

The subjects were classified according to their mean *J*-score on all 32 situations; this is referred to as "total *J*-score." This grouping was based on the assumption that the stronger the moral attitude of *S*, the lower his total *J*-score would be. Within each group,

Ss were then further divided according to sex. There are seven groups altogether, but only six for each sex, because one group had only males, and one only females.

Subjects

Ss were 29 male and 19 female students in an Introductory Psychology course given at Moravian College in the spring semester, 1963.

Results

Table 1 presents the results of the subjects-by-situations analysis. The data are presented separately for males and females, as well as for "professor leaves" (PL) and "professor stays" (PS).

TABLE 1
Justification for Cheating as a Function of Ss' Total J-score and Situational Anxiety and Hostility

		Professor Leaves						Professor Stays					
		Anxiety-Hostility Index						Anxiety-Hostility Index					
		0	1	2	3	4		0	1	2	3	4	
Total J-score	N	Number of Situations						Number of Situations					
		1	4	6	4	1	16	1	4	6	4	1	16
<i>Males</i>													
1.00	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
...	0
1.73	7	1.00	1.29	1.67	2.11	3.00	1.72	1.00	1.25	1.78	2.23	2.58	1.74
2.29	8	1.00	2.03	2.45	2.63	3.25	2.35	1.13	1.72	2.21	2.94	2.50	2.22
2.66	3	2.00	2.13	2.50	3.33	4.00	2.69	1.00	1.83	2.72	3.25	4.33	2.63
3.17	7	2.14	2.65	3.45	3.75	4.29	3.30	1.72	2.62	3.17	3.47	3.86	3.05
3.66	2	2.00	3.13	3.67	4.50	5.00	3.72	1.50	3.00	3.50	4.50	5.00	3.59
2.41	29	1.45	2.02	2.50	2.86	3.48	2.46	1.24	1.85	2.39	2.89	3.11	2.36
<i>Females</i>													
1.00	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.14	2	1.00	1.00	1.17	1.13	1.50	1.13	1.00	1.00	1.17	1.25	1.50	1.16
1.74	4	1.25	1.50	1.83	2.43	2.50	1.91	1.25	1.19	1.63	1.81	2.25	1.58
2.36	2	1.50	2.00	2.58	3.00	3.50	2.53	1.50	1.88	2.00	2.75	3.00	2.19
2.84	4	2.00	2.50	2.75	3.69	3.75	2.97	1.75	2.00	2.92	3.19	3.50	2.72
3.28	1	2.00	3.00	3.67	3.75	5.00	3.38	1.00	2.50	3.67	3.75	5.00	3.19
...	0
1.82	19	1.26	1.63	1.85	2.24	2.63	1.90	1.26	1.42	1.78	1.99	2.26	1.74

Each of the four sections of Table 1 forms an almost perfect "fan." Low J-values in the upper left-hand corners increase consist-

ently in both directions, until they reach a high point in the lower right-hand corners. For a given sub-group of *Ss*, at any one point on the total J-score continuum, J-values tend to increase as a function of how much anxiety and hostility are aroused by the situation descriptions. At the same time, for a given group of situations, at any one point on the A-H continuum, J-values tend to increase as a function of *Ss*' total J-scores.

Comparison of the means for males and females shows that the females had lower J-scores than the males at every point on the A-H continuum for PL, and at all but the 0-point for PS.

Comparison of the means for PL and PS shows that both males and females consider cheating more justified when PL. This is true in four of the five male and four of the five female groups in which such a comparison is meaningful; it is of course meaningless for those with total J-scores of 1.00.

Discussion

The observed "fans" suggests that it is possible to approach systematically yet concretely (Lewin, 1935) the subject-situation interactions in the measurement of attitudes. The "fans" are not forced by mathematical factors, although the use of total scores means that the final cell is determined when the others are given. There is no mathematical reason why *Ss* with low total J-scores and *Ss* with high total J-scores could not have reacted to the situations with differently shaped curves.

Furthermore, merely offering the subjects different situations does not automatically assure the observed pattern. To be sure, the presence of the 32 situation descriptions constitutes what Orne calls a "demand characteristic" (Orne, 1962); the subjects must have suspected that the situations were there for a purpose. The problem for a subject would be: what purpose? Will the professor-researchers think him a "good" subject if he says cheating is never justified, or if he says that it sometimes is? The obtained relationships between attitudes and situations are certainly contrary to the values of professors.

There is no empirical evidence about what *Ss* considered "being a good subject," except that they did answer the questions. However, the *Ss* did work under pressure of time, because they had so many questions to answer in one class hour, which means that any

conscious, calculated effort to "scale" answers would probably have been thwarted; Ss had to react to each situation quickly and as a whole. There were even a few complaints about "the fact" that the situations were all alike, and that answering the questions felt like guessing. Even without the time pressure it seems unlikely that Ss would have wanted to or been able to "analyze" the situations, and then scale their answers accordingly; however, the answers might then have been more censored, with lower J-values in every cell.

Not only is it possible to make subject-situation or subject-question interactions the focal point of attitude measurement, but it is important to do so if we wish to understand what responses to self-report inventories mean. The questions in such inventories are not always interpreted in a uniform manner by all respondents. It is sometimes argued that this need not be of concern in empirical research; it is sufficient if "neurotics" say "no" five times as frequently as "normals" do. This principle of validation has been used successfully in the SVIB and the MMPI. In theoretically oriented research, however, the goal is to understand such data, and such understanding involves theories and measuring instruments that are closely interrelated (Cronbach and Meehl, 1955). The situation-bound questionnaire used in this study was a theoretically derived instrument, based on situations hypothesized to be significant to the Ss. The importance of taking the situations or questions into account, in an effort to control to some extent what the subject has in mind when he reads each question, may be seen by considering the following example. Suppose the question had said only, "Do you consider cheating justified?" Suppose further that those with the highest total J-score rationalized their inclination to circle four or five by thinking about "those picky tests," and that those with lower total J-scores thought instead, "No, it isn't. Sometimes you run into a poor professor, and then it's tempting, but. . . ." In other words, the "high's" might think about A-H index 3 and 4, the "low's" about A-H index 0 and 1. If that happened, the experimenters wouldn't know it, but would then get as "individual differences" the entries on the upper-left to lower-right diagonal instead of the total J-scores. To be sure, the figures on the diagonal are data, yet they give a limited understanding of "individual differences." Under still different conditions, data might be collected

which represents the entries on the other diagonal (upper-right to lower-left), and in that case erroneous conclusions would be drawn, either about psychological functioning, or the measuring instruments, or both.

The subject-situation approach to attitude measurement has another advantage. Many attitudes are "undesirable," at least at one end of the continuum, and the problems this raises are well documented (Cronbach, 1960). Perhaps the situational questions can serve as a partial answer to problems created by conscious and unconscious "faking," since the situations provide built-in rationalizations: the student can bring himself to say that cheating is at least sometimes justified because "the situation is so bad." If this view is correct, it should follow that justification scores would be lower in a group asked only a general question about the justification for cheating than in a comparable group given a situational questionnaire.

Within the present study, the Ss appear to hold attitudes that they themselves might judge as "undesirable": cheating is justified in certain situations, and is not rejected point-blank as immoral. Not only is it condoned when the situation provokes anxiety and hostility, but in addition, it is considered more justified when the professor leaves the room than when he stays. When a professor leaves the room during a test, he does two things: he provides increased opportunity to get away with cheating, and he is putting students on their honor. In our sample, opportunity appears to win out over honor, which suggests that "justification" is more nearly a practical than a moral matter altogether. Using a more direct approach to such a distinction would be asking Ss to express a culturally and personally less accepted self-image.

In future studies, it will be important to classify the Ss on a sounder theoretical basis, rather than only by total score. This will be particularly pertinent in considering further the construct validity of such a questionnaire. As a first step in this direction, data were analyzed separately for males and females. The females were found to have lower J-scores than the males in almost all of the situations, a datum which was expected in view of differential sex roles and characteristics in our society. In a recent report, Burton suggests that girls may be more conforming cognitively to the general moral code than boys, though not necessarily behaviorally;

such greater cognitive conformity would be reflected in responses to questionnaires (Burton, 1963).

The one isolated datum involving a classification of subjects not based on the questionnaire is not sufficient, however. In the near future, data will be analyzed which was obtained from an entire freshman class. Subjects will be classified according to sex and planned-on major. One would expect, for example, that the pre-theology majors would verbalize "justification" least readily, but the shapes of the curves, again plotted separately for PL and PS, and against the A-H index, are harder to predict.

In extending the applications of this approach, it would not always be necessary to scale the situations, nor to base the situations exclusively on environmental elements. In a second questionnaire that was used with the freshman sample mentioned above, for example, reference was made to the Ss' needs or aspirations, using as one dichotomy, "To date, your grade in the course is D or less—To date, your grade in the course is B or more."

Questionnaires could easily be developed in quite different settings. For example, groups of adolescents (normal, psychopathic, schizoid) might be asked to indicate how tempted they would be to steal and/or how justified it would be to steal if a) they needed the money—didn't need it; b) the amount was \$5-\$50-\$100-\$500; c) they felt that nobody—parent(s)—police—peers would find out; etc. Or again, workers might be asked how they would feel about certain changes (about safety regulations or rest pauses, etc.) under various specific conditions in which they can readily imagine themselves. It would not always be possible to develop a combined index for all of the variables; the important thing would be to direct analyses at subject-situation interactions conceptualized on the basis of available psychological theory.

REFERENCES

- Allport, G. W. *Personality, a Psychological Interpretation*. New York: Holt, 1937.
- Burton, R. V. "Generality of Honesty Reconsidered." *Psychological Review*, LXX (1963), 481-499.
- Brunswik, E. *Systematic and Representative Design of Psychological Experiments*. Berkeley: University of California Press, 1947.
- Campbell, D. T. "Recommendations for APA Test Standards Re-

- garding Construct, Trait, and Discriminant Validity." *American Psychologist*, XV (1960), 546-553.
- Campbell, D. T. and Fiske, D. W. "Validation by the Multitrait—Multimethod Matrix." *Psychological Bulletin*, LVI (1959), 81-105.
- Cronbach, L. J. *Essentials of Psychological Testing* (2nd ed.), New York: Harper, 1960.
- Cronbach, L. J. and Meehl, P. E. "Construct Validity in Psychological Tests." *Psychological Bulletin*, LII (1955), 281-302.
- Hartshorne, H. and May, M. A. *Studies in Deceit*. New York: Macmillan, 1928.
- Lewin, K. *A Dynamic Theory of Personality*. New York: McGraw-Hill, 1935.
- Murphy, G. *Personality*. New York: Harper, 1947.
- Orne, M. T. "On the Social Psychology of the Psychological Experiment." *American Psychologist*, XVII (1962), 776-783.
- Steininger, M., Johnson, R., and Kirts, D. "Cheating on College Examinations as a Function of Situationally Aroused Anxiety and Hostility." *Journal of Educational Psychology*, LV (1964), 317-324.



THE DEVELOPMENT OF PERSONALITY FACTORS IN CHILDREN AND ADOLESCENTS¹

MICHAEL S. BLACK

Institute for Juvenile Research

RECENTLY, several investigators have attempted to study personality development from the structural point of view, by comparing personality structure at several ages (Cattell, 1947; Cattell and Coan, 1957; Cattell and Gruen, 1953; Peterson, 1960; Peterson and Cattell, 1959). The usual method has been to factor analyze behavior rating scales of children at various ages, and to compare the factors obtained with those obtained in previous studies, in order to determine what, if any, changes were taking place in personality structure as a function of age. These studies, for the most part, have led to the conclusion that the same personality factors are present at all ages examined.

Cattell (1947) factor analyzed peer ratings of college men in residence halls and identified 11 factors. Cattell and Gruen (1953) factor analyzed peer ratings of 10 to 14 year olds using primarily the same variables as Cattell (1947). They isolated six factors which they identified either as direct representatives or as com-

¹This paper is based upon a thesis submitted in partial fulfillment of the degree of Doctor of Philosophy in Psychology at the University of Illinois. The research was partially supported by USPH Predoctoral Research Fellowship No. MPM-18587. Many thanks are due to Donald R. Peterson and Ledyard R. Tucker, the supervisors of the thesis, for their generous assistance and encouragement. The author also wishes to thank Conan Edwards, Norman Gore, and Mark W. Bills, superintendents, respectively, of the Danville, Decatur, and Peoria, Illinois Public Schools and the University of Illinois' Child Development Laboratory for their kind assistance in obtaining subjects for this study.

binations of the 11 factors found in the former study. Cattell and Coan (1957) factor analyzed teacher ratings of 6-8 year olds, and visually compared the 11 factors obtained with those obtained in the Cattell (1947) study. They concluded that the factors were the same in both studies. Peterson and Cattell (1959) factor analyzed teacher ratings of 4-5 year olds, and isolated 12 factors, nine of which they quite positively identified with factors previously found among older subjects.

Claims of factor similarity in all these studies were supported almost entirely by "visual" matching of factor patterns. The latter technique amounts only to a judgment by the investigator as to the resemblance between one factor and another. When more precise indices of factor similarity are employed, the resemblance between allegedly similar factors is often far weaker than it first appears to be.

Peterson (1960) correlated the factor loadings obtained in the three studies of children cited above with those obtained in Cattell's (1947) study of adults. Mean correlations of .19 to .46 appeared for loadings on allegedly similar factors, depending on whether certain studies employing contaminated data and questionable analyses were omitted or included. Even the most favorable results, however, were unconvincing. Peterson concluded that correlation coefficients of this magnitude, even if significant, did not support the claim of previous authors that the factors obtained in the several studies were identical.

It was suggested, however, that most of the factors retained in the previous studies accounted for too small a proportion of the total variance to be of much practical significance. All but the two largest factors in each study were therefore eliminated, and columns of factors loadings were again correlated across the four studies. This time a mean correlation of .81 was obtained for apparently similar factors. Peterson concluded that this was large enough to support a contention of factor similarity. Thus, the two factors accounting for most of the common variance did seem to be stable over the age range examined while the many factors accounting for only a small proportion of the common variance did not.

A recent study by Linn (1964) has helped explain these findings. Random normal variates were intercorrelated and factored along with the "real" ratings employed in the Peterson and Cattell

(1959) study. When the number of factors exceeded two or three, the distributions of loadings for random variables became indistinguishable from those of the presumably "real" variables.

None of this, however, helps resolve the issue of possible changes in personality structure during childhood: and the problem seems important. If the time and manner of the development of various personality tendencies can be established with any degree of accuracy, this might offer a useful basis for the construction of a theory of personality development. Most empirical studies have appeared to show that the same factors are obtained at all ages, either a large number of narrow factors, or a small number of broad factors, depending upon the particular methods used. Such findings are contrary to generally accepted theory about personality development, and thus require careful examination. According to some widely accepted principles of personality growth, behavior becomes more highly organized and specific as age increases. That is, behavior progresses from diffuse activity in infancy to a high degree of organization and differentiation in adulthood (Ausubel, 1958; Berelson and Steiner, 1964; Jersild, 1954; Lewin, 1954; Witkin, et al., 1962). It seems reasonable to suppose that the development of personality structure might follow an analogous course; that is, from a single, general factor in infancy to a larger number of factors in adulthood. Other changes in personality structure might, of course, also occur.

To date, all attempts to examine personality factors across different age levels, have been done by comparison of factor patterns derived from totally independent studies. Consequently, the results have been confounded to an unknown extent by differences between the studies other than age of the population. The several studies have used, variously, teacher ratings, parent ratings, and peer ratings. The methods of rotation have varied from study to study, as well as the exact combination of variables used.

The technique of obtaining separate factor patterns at each age level, and subsequently matching these, also raises some problems. Although adequate measures of factor similarity are available (Harman, 1960), only two situations are well handled by this method: the case where the index is so high as to indicate identity of the two factors, and the case where the index is so low as to indicate independence of the two factors. Interpreting the relations

among a large number of factors, when all are interrelated to a moderate extent, would be, to say the least, ambiguous.

A more suitable technique has recently been developed by Tucker.² In this method, a single set of factors is obtained for all age levels. The variance accommodated by the factors at each age level, and the intercorrelations among the factors at each age level, are estimated. The variance of a factor is generally taken as an indication of the size, or importance, of that factor relative to other factors (Harman, 1960; Peterson, 1960). In the same manner, the variance, in Tucker's method, is used as an indication of the size, or importance, of a factor in one age group, relative to that in other age groups. To the extent that a factor is degenerating with age, its variance would be expected to decrease systematically with age; to the extent that a factor is emerging with age, its variance would be expected to increase systematically with age; to the extent that two factors are differentiating with age, their intercorrelation would be expected to decrease systematically with age; to the extent that two factors are merging with age, their intercorrelation would be expected to increase systematically with age. Discovery that variance is concentrated in a single factor for young children and diffused among many factors at later ages might assume some interest in the light of related principles of growth. A shift in variance concentration from one set of factors to another over the course of time could lead to the formation of new principles of personality development, with empirical specifications established from the beginning.

This study was designed to investigate changes that might occur in personality structure, as a function of age, using data collected in a uniform and systematic way at several different age levels, and employing a more powerful method of data analysis than any previously employed.

Method

Instruments

To permit articulation with previous findings and to facilitate obtaining the necessary large sample, ratings were used as the basic

² Tucker, L. Personal communication, 1962.

data medium. A rating schedule was composed of 51 eight-point bipolar scales. Comprehensive coverage of the personality domain was sought by employing variables over the entire personality sphere (Cattell, 1957). The 44 variables in the *Personality Sphere: Most Condensed Form* (Cattell, 1957) were used as a descriptive base. To these were added seven variables from the Fels Behavior Rating Scale (Richards and Simons, 1941), because of theoretical interest and special pertinence to children. The items were arranged so that each teacher could rate all his subjects on one variable at a time, in order to minimize the halo effect, logical error, and errors of proximity (Guilford, 1954).

Subjects and Procedure

The subjects were 700 students in the public school systems of Danville and Peoria, Illinois, and the Child Development Laboratory of the University of Illinois. In each of the two participating school systems, one teacher at each grade, from grades 1 through 12, was asked to rate the students in one of his classes on each of the 51 variables; that is, two raters were obtained for each grade, each representing a different school system. In the Child Development Laboratory, four teachers rated two classes each.

Seven grade levels, consisting of two grades each, were constituted; thus, the first level consisted of subjects of nursery school and kindergarten age, the second level consisted of first and second graders, the third level of third and fourth graders, and so fourth, to the seventh level which consisted of eleventh and twelfth graders.

All of the data for raters falling into any given grade level in each of the school systems were pooled. Thus, in each grade level in each school system, data from two raters were available, one rater from each of two grades represented. For Levels 2 through 7, 25 boys and 25 girls were chosen randomly from each of these groups to constitute the final sample to be analyzed. At Level 1, where only one school system was represented, 50 boys and 50 girls were chosen randomly from the available data. Thus, the final sample consisted of 700 subjects, 100 at each grade level. Within each grade level, the subjects were evenly divided between males and females, and, except at Level 1, between the two public school systems.

Analysis of Data

For the subjects in each grade level, a matrix of covariances among the ratings was obtained. The seven covariance matrices were added, and the results divided by seven, in order to obtain a mean covariance matrix.

Communalities, estimated by Tucker's method (Parker, 1963), were placed in the diagonal cells of the mean covariance matrix. The matrix was factor analyzed by the principal axis method. Factors were extracted until the total number accounted for 100 percent of the estimated communality, which resulted in the extraction of four factors.

Inspection of the eigenvalues indicated that the last large decrease in the size of the eigenvalues occurred between the fourth and fifth factors, the fifth and beyond having apparently reached a plateau, with only small, steady decreases. This, together with the fact that four factors accounted for 100.6 percent of the estimated communality, led to the conclusion that four factors were sufficient to account for the covariance among the original variables.

The four principal axis factors were then rotated to both the varimax criterion of orthogonal simple structure and the binormamin criterion of oblique simple structure.

Inspection of the plots of factor pairs of both the varimax and the binormamin rotations, indicated that the latter had approached simple structure considerably better than had the varimax rotation. The simple structure of the varimax factors was only fair, with many variables clustering between the axes, rather than on them. The simple structure of the binormamin rotation, on the other hand, appeared to be quite good. In addition, the binormamin factors made better sense than the varimax. For these reasons, it was decided to retain the binormamin solution, rather than the varimax.

A matrix of covariances and variances among factors, for each grade level, was obtained by means of Tucker's method for estimating factor variance.^{3, 4}

³ Tucker, L. Personal communication, 1962.

⁴ The computations for this study were carried out on the IBM 7094 digital computer at the University of Illinois' Digital Computer Laboratory, which is partially supported by the National Science Foundation Grant No. NSF-GP-700.

Results and Discussion

TABLE 1
Salient Variable Pattern of Binormamin Related Factors

Variable	Factor 1	Factor 2	Factor 3	Factor 4
1. Conscientious-Unconscientious	1.65			
2. Adaptable-Rigid	1.22			
3. Emotional-Calm		.64		.64
4. Conventional-Unconventional	1.12			
5. Not jealous-Prone to jealousy	1.19		-.64	
6. Considerate, polite-Inconsiderate, rude	1.51			
7. Determined, persevering-Quitting	1.62	.80		
8. Tender-Tough, hard	1.14			
9. Self effacing-Egotistical	1.13	-.64		-.59
10. Energetic, alert, active-Languid, fatigued, slow	1.02	1.15		
11. Assertive-Submissive		1.33		
12. Not attention seeking-Attention seeking		-1.38		
13. Interested in others-Cool, reserved	1.12	.72		
14. Socially, culturally, mature-Socially, culturally, immature	.83			
15. Thoughtful, pensive, original-Unthinking, goes vigorously with groups	1.31			
16. Resourceful-Not resourceful	.95			
17. Gregarious, sociable-Self contained	1.37	.80		
18. Patient, not demanding-Demanding, impatient		1.26		
19. Quiet, composed-Boisterous, noisy, rowdy	1.43	-.98		
20. Bold, adventurous-Cautious, retiring, shy, timid	1.18	-1.16		
21. Self reliant, independent-Dependent		1.49		
22. Happy go lucky-Anxious, careful, worrying	1.38			.49
23. Responsible, dependable-Irresponsible, undependable		.61	-.72	.63
24. Relaxed-Tense, high strung	1.87		.65	
25. Trusting-Suspicious	1.04			.65
26. Gay-Solemn	1.44			
27. Happy-Sad		1.17		
		.90		

TABLE 1—Continued.

28. Good natured, easy going-Irritable, spiteful	1.14		
29. Insistently orderly- Disorderly	1.22		.41
30. Tolerant of stress- Easily distressed	1.27		
31. Inquisitive, curious, interested-Uninquisitive, uninterested	.87	.98	
32. Cooperative-Obstructive	1.38		
33. Ready to admit mistakes-Reluctant to admit mistakes	1.47		
34. Mannerly, polished-Crude	1.49		
35. Not prone to daydream-Prone to daydream	1.16		
36. Obedient-Disobedient	1.34		.72
37. Unshakeable poise-Easily upset	.70		
38. Sensitively imaginative-Practical, logical	.95		
39. Esthetically fastidious-Lacking artistic taste	.72	.66	
40. Marked interested in opposite sex-Slight interest in opposite sex		1.23	
41. Frank, expressive-Secretive, reserved		1.53	
42. Talkative-Taciturn, silent, introspective	1.26		
43. Not sadistic-Destructive, sadistic	1.34		
44. Likes school or work-Dislikes school or work	1.02	1.19	
45. Competitive-Non competitive	1.10		.41
46. Kind-Cruel	1.21		.56
47. High degree of emotional control-Low degree of emotional control		1.19	
48. Active-Inactive	1.37	— .77	
49. Peaceful-Quarrelsome	.69	.82	
50. Sense of humor-No sense of humor	.89		
51. Suggestible-Negativistic			

Note—The signs of factor loadings and the positions of variable anchor terms have been reversed where the needs of clarity have required it.

The Factors

The first and most outstanding finding of the study was that a very limited number of very familiar factors emerged from the analysis. Recent suggestions that personality rating studies yield only a few factors (Linn, 1964; Peterson, 1960), rather than the many factors isolated in most earlier investigations (e.g., Cattell, 1947) have always been subject to doubt because of limitations of the methods employed. The more systematic procedures and more sophisticated analyses of the present research have apparently not altered the fact that raters perceive others (or at least that teachers perceive children), in a rather massive, poorly differentiated way.

The most striking facts about Factor 1 are the large number of variables that load upon it, and the highly evaluative connotation of nearly every variable. Bearing in mind the fact that these ratings are those by teachers of students, the two poles of the variables seem to represent just the sort of behavior that teachers would, and would not, respectively, like in their students. The behaviors represented could also be considered those of socially well-adjusted versus socially poorly-adjusted people. This factor could therefore be interpreted either as general adjustment, or as a dimension of evaluative semantic meaning, as defined by Osgood and others (Osgood, Suci, and Tannenbaum, 1957).

Factor 2 variables seem to have evaluative connotations to a much lesser, almost negligible extent. Instead, the common theme seems to be one of outward versus inward direction of activities, or Introversion-Extroversion. Interpretation as semantic "dynamism," i.e., a combination of perceived activity and potency (Osgood, *et al.*, 1957) is equally plausible.

Factors 3 and 4, although rotated with Factors 1 and 2, are of negligible interest. They accounted for only a small part of the total communality, the first two factors accounting for 88.45 percent of the common variance, and had only a few barely salient loadings. While they may be "significant" in a mathematical sense, they seem to contribute nothing to a better understanding of personality structure, and it would be more parsimonious to omit them.

The four factors obtained do not appear to correspond to any of the factors found in Cattell (1947), Cattell and Coan (1957), Cat-

tell and Gruen (1953), or Peterson and Cattell (1959). They are, of course, considerably fewer in number, and appear to cut across those found in the above studies.

On the other hand, the two largest factors (General Adjustment and Introversion-Extroversion) do appear to be very closely similar to the factors which emerged in Peterson's (1960) reanalysis of data by Cattell and his colleagues. These also clearly resemble the second-order factors Becker (1960) derived from teacher ratings of kindergarten children. Eysenck, of course, has obtained factors in a number of studies (Eysenck, 1947, 1953) which resemble the General Adjustment (as Neuroticism) and Introversion-Extroversion dimensions of the present study. Since Eysenck used a different set of variables, no exact comparison is possible, but the interpretation of Eysenck's factors (Neuroticism and Introversion-Extroversion), seems to be quite similar to the two present factors of General Adjustment and Introversion-Extroversion.

The reason for the discrepancies between the Cattell, *et al.* (1953, 1957, 1959) studies, on the one hand, and the Eysenck (1947, 1953), Peterson (1960), and the present studies, on the other hand, seems to be a matter of the number of factors rotated. Peterson (1960, 1965), presents good reason for favoring a small number of factors, not the least of which is the likelihood that many of the small factors in previous research may be chance phenomena. Considerations of variance and descriptive efficiency have also been dealt with in previous analyses (e.g., Peterson, 1960). Extracting large numbers of low variance factors defeats the purpose of factor analysis; that is, the achievement of descriptive parsimony. Two additional criteria were available in this study: 1) factors beyond the second appeared to be meaningless, and 2) four factors accounted for 100 percent of the estimated communality. Under the circumstances, there appeared to be no justification for extracting more factors.

Variances and Interrelations

Since the sampling error of variances estimated by the Tucker method is unknown, no exact test for homogeneity of the variances between the age groups is available. An approximate test was, however, applied to each factor. For each factor, a few relatively pure variables loading on that factor were selected. At each grade

TABLE 2
Factor Variance Estimated by Tucker's Method

Grades	Factor Number	
	1	2
NS-K	.78	1.67
1-2	.75	.72
3-4	1.02	.81
5-6	1.68	1.26
7-8	.87	.68
9-10	.98	.62
11-12	.92	1.23
Mean	1.00	1.00

Note—The mean variance of each factor has been scaled to 1.00.

TABLE 3
Factor Correlations Estimated by Tucker's Method

Grades	Correlation
NS-K	-.15
1-2	-.18
3-4	.11
5-6	.01
7-8	.09
9-10	.13
11-12	.18
Mean	.02

Note—The mean correlation was computed by obtaining the square root of the mean squared correlation.

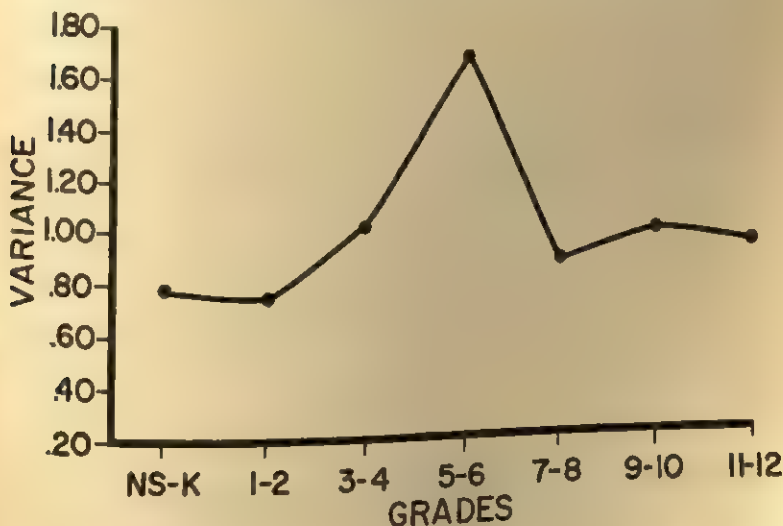


Fig. 1. Variance of each grade level of Factor 1.

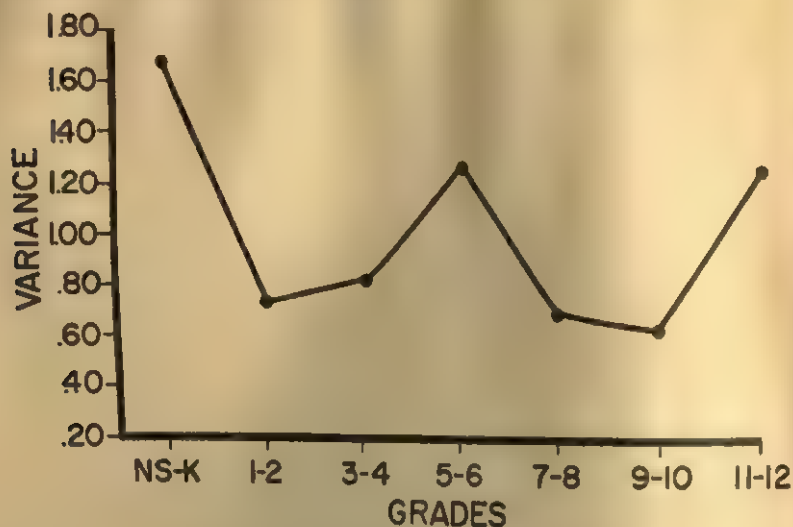


Fig. 2. Variance of each grade level of Factor 2.

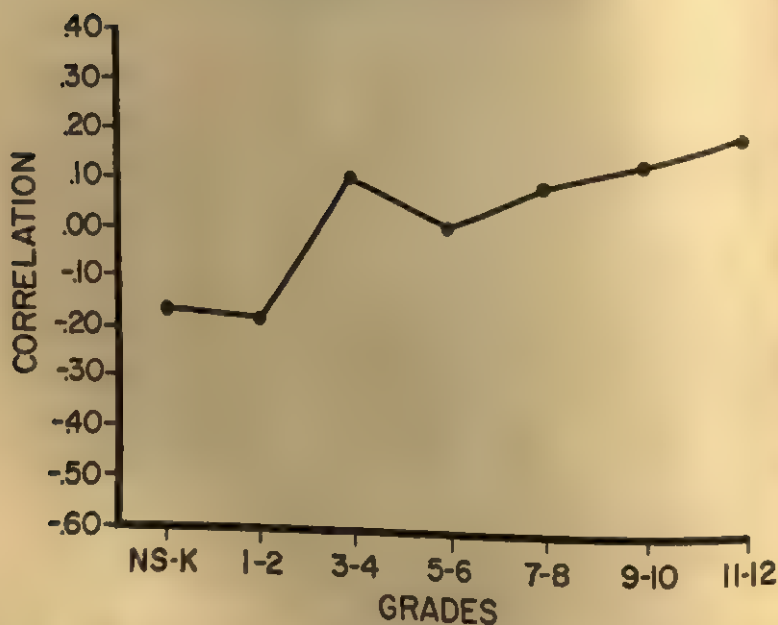


Fig. 3. Correlation of Factors 1 and 2 at each grade level.

level, the variance of each variable was computed, and the variances summed across the several variables selected. Hartley's Test (Walker and Lev, 1953) was then applied to the summed variances.

The test is approximate, since 1) no variables are completely pure, and 2) the test involves only a few variables, whereas the Tucker estimated variances are based on all variables. The variances of the several age levels were significantly different, at a significance level of $<.01$, for both factors.

The reliability of the behavior indicants used in the present study should also be taken into account. Since several raters were pooled at each age level, the effect of differences between rater means must be considered as a possible source of confounding. If such differences do exist, they would tend to inflate the variance estimates obtained by Tucker's method; that is, the total variance at any grade level would be in part intrarater variance and in part interrater variance (Hays, 1963). It is, of course, the intrarater variance which is desired, as a measure of the variability among subjects. Further, if the interrater variance is inconsistent from one grade level to the others, it would have a differential effect on the variance of the several grade levels, thus contaminating the results.

Hays (1963) proposes a method for estimating the true interrater variance. This was done for each of the seven grade levels, on each of the two meaningful factors. For each factor, two variables of reasonably pure factor loadings were selected, the scores summed, and the intrarater and interrater variance components estimated. The interrater variance was negligible at all grade levels on Factor 2. On Factor 1, the interrater variance was quite small, the mean interrater variance being only 4 percent of the total variance. It was also quite consistent among the grade levels. This would have no critical effect on the differences between the variances of the seven levels, except to make the Hartley test slightly more conservative, since the variance ratios are slightly decreased.

Interrater correlations were not obtained in this study, though, of course, such formation is always useful to have. Aside from the necessities imposed by that fact each class had only one teacher who knew the students well enough to give useful information about them, reliability data of this type did not seem necessary; rather, estimates of variance components, as mentioned above, seemed more pertinent. Previous studies have shown the interrater correlations for ratings of the kind used in this study, to hover in the neighborhood of .50 and .70 (Peterson and Cattell, 1959). Factor score reliabilities are usually a little better (Becker, 1960). The mean com-

munality of all the variables was .73 derived from the estimated communalities, and .75 derived from the communalities computed with the factor solution. Since the communalities, which are usually considered as lower bounds of reliabilities (Harman, 1960), were reasonably high in the present study, and since the factors which emerged were so clearly compatible with those from previous studies, reliabilities must have been adequate.

Interpretation of the trends of variance and intercorrelations is a more difficult matter. Inspection of Tables 2 and 3, along with a look at the corresponding figures, at first yields a general impression of disorder, and generates a suspicion that the results might be due to chance, despite the approximate test for heterogeneity of variance which was applied. The variance trends certainly are not linearly related to grade level. However, it may be profitable to discuss some of the trends in variance and intercorrelations to see what sense can be made of them.

For Factor 1, there appears to be a trend of increasing variance, beginning at the nursery school-kindergarten (NS-K) group, reaching a peak at the 5th and 6th grades, followed by a sharp drop to the near-mean level, where it remains from grades 7-12. In other words, children appear to be becoming more diverse in their general adjustment, (or, more properly, teachers' evaluation of general adjustment) until grades 5-6, followed by greater similarity in grades 7-12. To the extent that the ratings actually reflect child behavior, this trend might be related to the difficulties associated with early adolescence, and the different ages at which children reach puberty. Before the latest years of elementary school, all children are pre-adolescent. Some time later, all have reached sexual maturity. But around the sixth grade, some children are entering adolescence and others are not. Greater variance might therefore be expected at this time.

Factor 2 shows three peaks of high variance among subjects: at the nursery school-kindergarten level, the 5-6 grade level, and the 11-12 grade level. These three peaks seem to coincide with periods of transition in the lives of children and adolescents. The first marks the child's transition from being continually at home to the school situation, where he is expected to interact with his peers to a much greater extent than previously. At such a time, great diversity in introversion and extroversion might be expected, with

some children "getting into the swing of things" quickly, and others remaining reticent for quite a while. The second peak occurs, as with Factor 1, at the 5-6 grade level, and may also be related to the onset of puberty. The third peak occurs during the final years of high school. It is at this time that adolescents are beginning another transition in their lives: from high school into the adult world, or into college.

Factors 1 and 2 have a slight negative correlation at grades NS-2, then show a trend toward an increasingly positive correlation, terminating with a slight positive correlation at grades 11-12. In other words, good adjustment (or high evaluation) seems at first to be slightly related to introversion, but the direction then reverses, as age increases, and good adjustment ends up being slightly related to extroversion. In view of the fact that, in the U. S. culture, extroversion is highly valued in adults, but not so much in very young children, this finding is not surprising. On the other hand, the relation is very slight. The correlation including all grades is .02.

A number of possibilities need to be considered as explanations for the failure to demonstrate convincing and clearly interpretable changes in the factor structure of personality as a function of age. The age range covered by this study was approximately 3 to 18 years; yet, it is quite conceivable that the important changes in personality structure occur before the age of 3 years. Although two factors are present at the lowest age level of this study, it is difficult to imagine describing the "personality" of a new-born infant along even that many dimensions. Some students of infant behavior have, in fact, emphasized a single dimension, such as excitability, or operant activity level. If only one dimension is appropriate to characterize infant behavior, and the two dimensions found in this study do emerge from a single, general, dimension, the 0 to 3 age range may be the proper one to investigate. Much of the literature has emphasized the importance of the first few years for the development of personality. (Ausubel, 1958; Fenichel, 1945; Hunt and Luria, 1958).

Further, there is no reason to suppose that the development of personality factors in the age range studied must be a linear function of age. It is just as reasonable, perhaps even more so, that such development might be the result of periodic biologic and psycho-

logic stresses to which the individual is subjected. This, indeed, is the implication of the patterns of factor variance obtained for Factors 1 and 2, with their peak variances always occurring at age levels when most individuals are at a transitional, and presumably stressful, period of their lives.

More important considerations, however, concern the meaning of verbally-defined measures of personality and the adequacy of ratings as measures of personality. Peterson (1965) has pointed out the very likely possibility that personality factors derived from ratings of verbally-defined traits have more to do with the perceptual and judgmental tendencies of the raters than with actual behavior of the subjects. If this is the case, then such "personality structure" would be expected to be reasonably similar at all age levels, and may bear little resemblance to the structure that might be obtained from, for example, observations of specific behavior acts, or objective measures of personality.

In view of the preceding arguments, generalization from this study must be sharply limited. "Personality structure" in the present context refers to "personality structure as derived from factor analysis of teacher ratings of verbally-defined personality traits, in children from nursery school through twelfth grade." Such a limitation is cumbersome, and restrictive, but it appears to be necessary.

Future research in this area might take the following recommendations into account.

- 1) The study should be replicated, extending the age range down to birth, for reasons mentioned previously, and up through adult levels, to determine whether the factor variances and intercorrelations continue to fluctuate, or level off; 2) if ratings are to be used at all, parent ratings, self-ratings, and peer ratings, as well as teacher ratings, should be obtained; 3) most important, variables should be used which are related more directly to observable behavior, in view of the present uncertainty about the meaning of verbally-defined personality factors. Generally, the question of changes in personality structure as a function of age needs to be systematically examined across more media and more situations than were included in the present study.

The problem of developmental change in personality structure is an important one. Analytical techniques for solution of the problem are now available. The major difficulty in specifying structural

change is now seen to reside in the character of the data which previous investigators, as well as the present one, have employed. Ratings may still be useful, but they should not be interpreted, simply and naively, as direct indicants of behavior.

Summary

This study was designed to investigate the factor structure of personality as a function of age. Teacher ratings were obtained of 700 grade and high school students (100 students in each of 7 grade levels) on 51 scales of personality traits. The trait ratings were intercorrelated and factor analyzed. Two meaningful factors were found. The variance contributed to each factor by each grade level, and the intercorrelations among factors at each grade level, were obtained and graphed.

The following results emerged.

1. Factors of perceived General Adjustment and Introversion-Extroversion emerged. Retention of a greater number of factors was not justified by these data. The results were in clear agreement with previous findings from personality rating research (Eysenck, 1947, 1953; Peterson, 1960; Becker, 1960).
2. No evidence of systematic changes in the factor structure of personality, as a function of age, was found. The generality of this finding is, however, limited to the particular type of variable, type of rater, and the age range of subjects used in this study.
3. Some age differences in factor variance and factor intercorrelations were found. These appear to be significant on the basis of an approximate test, but trends were difficult to interpret, and with a few exceptions bore no systematic relation to age. The possible relation of these trends to periods of transition and stress, as an alternative to linear relation with age, was discussed.

Limits of the study were considered, and recommendations for further research were stated.

REFERENCES

- Ausubel, D. P. *Theory and Problems of Child Development*. New York: Grune and Stratton, 1958.
- Becker, W. C. "The Relationship of Factors in Parental Ratings of Self and Each Other to the Behavior of Kindergarten Chil-

- dren as Rated by Mothers, Fathers, and Teachers." *Journal of Consulting Psychology*, XXIV (1960) 507-527.
- Berlson, B. and Steiner, G. *Human Behavior: An Inventory of Scientific Findings*. New York: Harcourt, Brace, & World, 1964.
- Cattell, R. B. "Confirmation and Clarification of Primary Personality Factors." *Psychometrika*, XII (1947), 197-220.
- Cattell, R. B. *Personality and Motivation Structure and Measurement*. Yonkers-on-Hudson, New York: World Book, 1957.
- Cattell, R. B. and Coan, R. "Child Personality Structure." *Journal of Clinical Psychology*, XIII (1957), 315-327.
- Cattell, R. B. and Gruen, W. "The Personality Factor Structure of 11-year Old Children." *Journal of Clinical Psychology*, IX (1953), 256-266.
- Eysenck, H. J. *Dimensions of Personality*. London: Routledge and Kegan Paul, 1947.
- Eysenck, H. J. *The Structure of Human Personality*. London: Routledge and Kegan Paul, 1953.
- Fenichel, O. *The Psychoanalytic Theory of Neurosis*. New York: Norton, 1945.
- Guilford, J. P. *Psychometric Methods*. New York: McGraw-Hill, 1954.
- Harman, H. H. *Modern Factor Analysis*. Chicago: University of Chicago Press, 1960.
- Hays, W. L. *Statistics for Psychologists*. New York: Holt, Rinehart, and Winston, 1963.
- Hunt, J. McV. and Luria, Zella. "Psychosexual Development: The Infant Disciplines." Unpublished manuscript, University of Illinois Library, 1958.
- Jersild, A. T. "Emotional Development." In L. Carmichael (Ed.), *Manual of Child Psychology*. New York: Chapman and Hall, 1954.
- Lewin, K. "Behavior and Development as a Function of the Total Situation." In L. Carmichael (Ed.), *Manual of Child Psychology*. New York: Chapman and Hall, 1954.
- Linn, R. L. "Use of Random Normal Deviates to Determine the Number of Factors to Extract in Factor Analysis." Unpublished M. A. thesis, University of Illinois, 1964.
- Osgood, C. E., Suci, G. G., and Tannebaum, P. H. *The Measurement of Meaning*, Urbana, Illinois: University of Illinois Press, 1957.
- Parker, J. L. "Communality Estimation." University of Illinois, Statistical Service Unit Library, unpublished manuscript, 1963.
- Peterson, D. R. "Age Generality of Personality Factors Derived from Ratings." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XX (1960), 461-474.
- Peterson, D. R. "The Scope and Generality of Verbally Defined Personality Factors." *Psychological Review*, LXXII (1965), 48-59.
- Peterson, D. R. and Cattell, R. B. "Personality Factors in Nursery

- School Children as Derived From Teachers Ratings." *Journal of Consulting Psychology*, XXIII (1959), 562.
- Richards, T. W. and Simons, M. P. "The Fels Child Behavior Scales." *Genetic Psychology Monographs*, XXIV (1941), 259-309.
- Walker, Helen and Lev, J. *Statistical Inference*. New York: Holt, Rinehart, and Winston, 1953.
- Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., and Karp, S. S. *Psychological Differentiation: Studies in Development*. New York: Wiley, 1962.



CHILD BEHAVIOR RATINGS: FURTHER EVIDENCE OF A MULTIPLE-FACTOR MODEL OF CHILD PERSONALITY¹

JOHN M. DIGMAN
University of Hawaii

FACTOR-ANALYTIC studies of child personality organization may be readily classified into two principal types of solution: simple (in terms of two or three factors) or complex (in terms of eight or more factors). Examples of the more parsimonious type of analysis are studies by Richards and Simons (1941), Peterson (1960), and Schaefer (1961). More complex are the solutions arrived at by Cattell and Coan (1957), Cattell and Gruen (1953), Stott (1962), and Digman (1963a). These quite different approaches to the problem of finding a meaningful factor model for child personality are paralleled by studies of adult personality structure: Eysenck's (1953) solutions have usually settled for two or three factors, while those of Cattell (1957) are usually in terms of a minimum of twelve.

A review of a number of factor studies of child personality (Digman, 1963b) suggested two reasons for the simpler solutions: inadequate facilities for a complete solution (e.g., the study by Richards and Simons, done long before the days of electronic assistance), and inadequate sample (it is not uncommon to find studies in which $N = 45$ for 30 variables, and studies in which the number of variables exceeded the size of the sample are not unknown).

¹ The preliminary factor analysis used the services of the Western Data Processing Center and BIMD-17, a program which terminates in a Varimax solution. The final solution used the services of the University of Hawaii's Computing Center.

A resolution of the simplicity vs. complexity issue has been suggested by Digman (1963b): the "two-factor" solutions are simply approximations of the first two second-order factors in this domain. The argument as to whether there are many or few factors depends, then, upon the level of abstraction at which one wishes to operate. At the primary factor level, there may be eight or more factors; at the second-order level there are probably substantially fewer. (The issue here is quite similar to the older question of "How many factors?" in the realm of intellect. Spearman's single-factor model and Thurstone's multiple-factor model are both "correct," in that they may be obtained from the same data. However, the view of intelligence as a simple, unitary construct appears valid only at a fairly high level of abstraction.)

The "two-factor model" of child personality appears to have been originally proposed by Peterson (1960), following his re-analysis of the data involved in a previous study by Peterson and Cattell (1958). Since many of the simpler solutions have been forced into parsimony because of small sample size or inadequate facilities for data processing, Peterson's solution appears as something of an anomaly: examination of the unpublished account of the original study² indicates that the data were very carefully gathered, the sample size was adequate ($N = 80$ for 36 variables), and the facilities of ILLIAC were employed in the analysis. In addition, the judges who rated the children were well qualified and appear to have given the investigators quite reliable data, as evidenced by an average inter-judge reliability of .70 for the 36 scales. Further, ranks obtained from the judges were converted to standard scores for each judge, and these in turn were averaged for the (usually three) judges involved before correlations were computed.

Examination of the unpublished details of the study have convinced the writer that this study constitutes some of the best prime data in the domain of child personality structure. In particular, since the data were unusually well gathered, they should throw light on the "simplicity-complexity" issue, and provide a good test for the generality of a multiple-factor model.

² The reader of the unpublished Peterson-Cattell manuscript is advised that it contains a number of clerical errors, including three serious ones in the correlation matrix. These were corrected for the analysis reported here.

Re-analysis of the Peterson-Cattell Data

Beginning with the correlation matrix, a factor analysis was conducted, following procedure comparable to the author's past studies (Digman, 1963a, 1963b). The number of useful common factors was estimated to be *eight* on the basis of two considerations: (a) the number of latent roots for the full correlation matrix in excess of 1.00; (b) consideration of the likely factor content of the scales in light of other studies. For the full matrix, seven roots were in excess of 1.00. Another analysis, with more variables and subjects (Digman, 1963b) has suggested ten factors, one of which (a "Tension" factor) was rather minor in the scheme of things. Since the Peterson-Cattell matrix was founded on fewer subjects and variables, it would be likely to support fewer common factors. One minor, but well established factor ("Sensitive Imagination"—very likely related to creativity) had no apparent representatives in the Peterson-Cattell list of variables.

Since the "rule of 1.00" suggested seven common factors, and other studies implied the possibility of eight, the analysis was conducted in terms of eight common factors. Communalities were conservatively estimated as the square of the highest correlation in the column. Following Varimax transformation of a principal-axes solution, an oblique transformation was undertaken, using the author's version of Cattell's Rotoplot technique, adapted for use on the University of Hawaii's IBM 7040 system.

Table 1 represents the final solution in terms of the reference vector matrix. Inasmuch as the width of the hyperplane was taken as $\pm .20$, elements in the matrix between these values have been supplanted with an asterisk for clarity.

Interpretation of the Factors

Factor I is clearly the Friendly Disposition vs. Hostility dimension of other analyses (Digman, 1963b). Characteristic 8 is the most salient here, and it is worthwhile to reproduce it in its complete form:

Tender	vs.	Tough
Governed by sentiment. Intuitive, empathic, sympathetic. Sensitive to the feelings of others. Cannot do things if they offend his feelings.		Governed by fact and necessity rather than by sentiment. Unsympathetic. Does not mind upsetting others if that is what has to be done.

TABLE 1

Oblique Reference Vector Matrix(Decimals omitted; items within $\pm .20$ indicated by *)

Scale (in abbreviated form)	I	II	III	IV	V	VI	VII	VIII
8. Tough, hard vs. tender	44	*	*	*	*	*	*	*
6. Inconsiderate, rude vs. considerate, polite	32	*	*	*	*	*	*	*
3. Unconscious vs. conscientious	31	*	*	*	*	*	*	*
28. Irritable, spiteful vs. good-natured, easygoing	27	*	*	*	*	*	*	*
33. Reluctant to admit mistakes vs. ready to admit mistakes	27	*	*	*	*	*	*	*
19. Boisterous, noisy, rowdy vs. quiet, composed	*	46	*	*	*	*	*	*
31. Inquisitive, curious, interested vs. uninquisitive, uninterested	*	48	*	*	*	*	*	*
10. Energetic, alert, active vs. languid, slow	*	36	*	*	*	*	*	*
24. Tense, high strung vs. relaxed	*	36	*	*	*	*	*	*
35. Not prone to day dream vs. prone to daydream	*	*	39	*	*	*	*	*
11. Assertive vs. submissive	*	*	29	*	*	*	*	*
9. Egotistical vs. self-effacing	*	*	27	*	*	*	*	*
22. Anxious, careful, worrying vs. happy-go-lucky	*	*	*	41	*	*	*	*

TABLE 1 (Continued)

Scale (in abbreviated form)	I	II	III	IV	Factor V	VI	VII	VIII
27. Sad vs. happy	*	*	*	34	*	*	*	*
20. Cautious, retiring, timid vs. bold, adventurous	*	*	*	22	*	*	*	*
26. Solemn vs. gay	*	-37	*	39	*	*	*	*
7. Determined, persevering vs. quitting	*	*	*	*	43	*	*	*
23. Responsible, dependable vs. irresponsible, undependable	*	*	*	*	34	*	*	*
16. Resourceful vs. non-resourceful	*	*	*	*	28	*	*	*
4. Conventional vs. unconventional, eccentric	*	*	*	*	*	52	*	*
17. Gregarious, sociable vs. self-contained	*	*	*	*	*	45	*	*
15. Unthinking, goes vigorously with group vs. thoughtful, pensive, original	*	*	*	*	*	40	*	*
13. Interested in others vs. cool, reserved	*	*	24	*	*	28	*	*
12. Attention seeking vs. not attention seeking	*	*	*	*	*	*	48	36
18. Demanding, impatient vs. patient, not demanding	*	*	*	*	*	*	43	*
30. Easily distressed vs. tolerant of stress	*	*	-22	*	*	*	43	*

TABLE 1 (Continued)

21.	Dependent vs. self-reliant, independent
5.	Prono to jealousy vs. not jealous
2.	Emotional vs. not emotional
25.	Suspicious vs. trusting
14.	Socially, culturally mature vs. socially, culturally immature
29.	Insisiently orderly vs. disorderly
34.	Mannerly, polished vs. crude
32.	Cooperative vs. obstructive
36.	Disobedient ¹ vs. obedient
1.	Rigid ¹ vs. flexible

¹Variables with all loading between $\pm .20$. Listed under factor for which variable has highest loading.

Readers familiar with Cattell's (1957) system would probably identify this factor as Schizothymia-Cyclothymia (A), while to Adler (1939) the factor would doubtless have suggested "Social Interest." Regardless of title, it appears to be a dimension of interpersonal affect, readily found in studies from the nursery school level (as in the present case) to the adult level.

Factor II has been interpreted as Cattell's Excitement factor (D). Cattell has commented on the "elusive" character of this factor, which has appeared in about one-half of factor studies concerned with rating data. The fact that it appears here, and is clearly evident in a re-analysis of the Richards and Simons data (Digman, 1963b), which is also based on nursery school observations, but is often not found at later ages, is entirely in accord with socialization theory, which would predict a compression of variability here. Another possibility, however, and one which is equally plausible, is that the nursery school situation provides for observation of subjects under many roles and circumstances, whereas observation for older subjects is more "specialized" (e.g., the teacher observer, the clinician observer, the supervisor observer, etc.).

Factor III, with its markers of "Egotistical" and "Assertive," variables which are salient for Cattell's *Dominance-Submissiveness* (E) factor, has accordingly been given that title.

Factor IV could be identified as either *Surgency* (Cattell's F) or *Ego Strength vs. Neuroticism* (Cattell's C). It is unfortunate that the collection of characteristics employed in the study did not include other indices of neurotic trend, such as defense mechanisms. The writer is inclined to the view that Factor IV represents the "misery" of the neurotic (Dollard and Miller, 1950), and this interpretation is supported by the position of this factor in the second-order space (see below).

Factor V, *Competence vs. Lack of Competence*, is clearly Cattell's *Superego Strength* (G) factor. The writer suggests the former title, however, if only for staying out of arguments with clinical associates! This particular dimension, incidentally, happens to be very clear in all factor studies of child personality which the writer has conducted or examined. It is possible, of course, to construct a list of characteristics for a factor study which would not include items relevant to this factor; in reality it is unlikely in the extreme that any sizable collection of behavior characteristics assembled by

anyone with even superficial knowledge of children would fail to include them.

Factor VI, with its high loadings on "Conventional," "Gregarious," and "Unthinking, goes vigorously with the group," is obviously a dimension of interest in group activities. It is very similar to a *Social Confidence vs. Social Diffidence* factor of other studies (Digman, 1963a, 1963b), and has been given this title. In Cattell's system, the factor would be identified as *Parmia vs. Threctia* (H).

Factor VII has been interpreted as *Relational Security vs. Relational Insecurity*. It is a factor easily isolated in child behavior studies (Digman, 1963b) and appears to reflect feelings of insecurity with respect to persons with whom there are basic affectional ties. Cattell has interpreted this collection of behaviors as *Protected Emotional Sensitivity (Premsia)*. It is entirely possible, of course, that the pattern might arise either from parental rejection (the "relational insecurity" interpretation) or from overprotection (the "protected emotional sensitivity" view).

Factor VIII is easily identified as a dimension of *Compliance vs. Non-compliance* (cf. Cattell's *Comention vs. Abcultion* (K) factor). Characteristic 12, Attention Seeking, seems at first glance to have an inappropriate sign here, but it is worth recalling that the data reflect the behavior of four-year-olds. That the socially precocious four-year-old enjoys the limelight is probably not news to observers of behavior at this age level.

Second-Order Analysis

Correlations between factors were computed by the relation

$$C_i = [(T')^{-1}D]' [(T')^{-1}D],$$

where T is the transformation matrix used to derive the oblique reference vector matrix from the preliminary Varimax matrix, and D is a normalizing diagonal matrix. Three factors were indicated by the "rule of latent root of 1.00." Residuals after extraction of the third factor were negligible (all within $\pm .08$). On the other hand, a two-factor solution here would have left several substantial residuals (e.g., .29, .23). The principal-axes matrix was then transformed by Varimax, and this was followed by plotting and single-plane shifts, preserving orthogonality. The resulting three-factor solution is presented in Table 2.

TABLE 2

Second-Order Factors
(Decimals Omitted)

Primary	Factor		
	A	B	C
Relational Security vs. Relational Insecurity	78	26	-20
Excitement vs. Apathy	-44	48	36
Dominance vs. Submissiveness	-07	69	38
Social Confidence vs. Social Diffidence	02	54	-35
Ego Strength vs. Neuroticism	38	-12	-55
Compliance vs. Non-compliance	68	-07	-21
Friendly Disposition vs. Hostility	80	10	-37
Competence vs. Lack of Competence	80	02	23

As in the case of previous second-order analyses (Digman, 1963a, 1963b), the first second-order factor is implied by *Friendly Disposition vs. Hostility*, *Compliance vs. Non-compliance*, *Competence vs. Lack of Competence*, *Relational Security vs. Relational Insecurity*, and, to a lesser extent, *Excitement vs. Apathy* and *Ego Strength vs. Neuroticism*. For this factor the label *Successful vs. Unsuccessful Socialization* has been offered in the other studies and seems quite appropriate here.

Extraversion vs. Introversion might be considered for the second factor, concerned as it is with *Excitement vs. Apathy*, *Dominance vs. Submissiveness*, and *Social Confidence vs. Social Diffidence*. However, the original meaning of the term, as proposed by Jung, would have to be modified to make the title appropriate here. The writer proposes *Freedom of Movement vs. Restraint* for this collection of primaries. The primary factor of *Sensitive Imagination*, missing in this analysis (because of failure to include appropriate indices), is typically found subsumed under this second-order factor (i.e., the *Sensitive Imagination* end of the factor is related to the *Freedom of Movement* end of the second-order factor). Since the *Sensitive Imagination* factor is probably "creativity," it is not surprising to find it associated with "Freedom of Movement." If, on the other hand, "Extraversion-Introversion" were retained as the title, the statement that creativity is indicative of extraversion would doubtless prompt any number of protestations and arguments.

The final second-order factor is chiefly concerned with *Neuroticism vs. Ego Strength*. To a lesser degree, it reflects *Hostility vs.*

Friendly Disposition, Excitement vs. Apathy and Dominance vs. Submissiveness. This is evidently an emotional factor, for which the writer has elsewhere (Digman, 1963b) proposed the label *Anxiety*. However, a broader concept, which reflects the common component of emotion evident in the primaries, is *Emotionality*. This third factor, by whatever title, is evident in two other, independent studies (Digman, 1963b; Walker, 1962). The latter study reports only primary cluster correlations, which have been factored by the writer (Digman, 1963b), giving an approximation to the second-order structure.

Contrast of the Present Solution with the Original, "Two-Factor" Solution

The eight-factor solution for the primary factor analysis is unusually clear-cut. The vast majority of variables have factorial simplicity, loading appreciably but one factor. A few (six of the 36) load two factors. Two variables appear to be factorially complex.

A two-factor solution, on the other hand (Peterson, 1960), results in many variables displaying factorial complexity (about one-half, depending on where the primary axes are placed).

On more theoretical grounds, the multiple-factor solution offers a concept of Hostility as distinct from (but related to) Neuroticism, whereas the two-factor solution offers the single concept of "General Maladjustment." It is worth noting that, were factor scores to be computed, the two-factor solution would very likely give an extremely hostile, but quite competent and superficially compliant child a relatively high score for "Adjustment." The multiple-factor model, of course, would give the child a relatively high score on "Socialization"; on the other hand, it would also present the pattern of primary scores underlying this socialization index. (Although different statistical principles are involved, the point of the above argument is somewhat akin to an interpretation of interaction in a complex analysis of variance: with substantial interaction, one may be able to make conclusions regarding a main effect; such conclusions, however, are rather on the crude side, and a more detailed exposition of effects is generally held to be more meaningful.)

It might be claimed that, while the two-factor solution is an over-

simplification of reality, rating data will not support more than two factors with reliability across studies. The fact of the matter is that any carefully done study in this domain will generally support eight to ten factors, and these factors can be easily identified, either by their similarity of content with other studies at the child level or the adult level (Digman, 1963b). Failure to achieve a satisfactory solution in some previous studies can be attributed to smallness of sample (the Richards and Simons study, for example, used but 40 subjects), overextraction (e.g., the Cattell-Coan study [1957] in terms of 13 factors), inadequate transformation (i.e., rotation), or careless data gathering. The data on which the present analysis is based were gathered with extreme care. Following an orthodox factor analysis with very conservatively estimated communalities, and with attention to a commonly used rule-of-thumb with respect to number of common factors, a solution was achieved which, by Thurstone's (1947) criterion of simple structure, is outstandingly clear-cut, both with regard to structure and interpretation.

Some Implications for Child Personality Research

Behavior ratings have been in disrepute for some time, for two reasons chiefly: (a) they often reflect numerous biases on the part of raters, and (b) there is widespread belief that ratings give little beyond "general, overall impressions" (halo effect). However, as in the case of any kind of data, these difficulties may be minimized by taking proper precautions in data gathering. Biases can be substantially reduced by proper choice of scaling method: e.g., the method of rank order, rather than single stimuli. In addition, as Guilford (1954) points out, the use of well trained and well instructed raters can do much to reduce the "subjectivity" of ratings. "Halo" may be minimized by separating judgments in time, by judging all persons on one scale at a time (as was done in the Peterson-Cattell investigation), and by providing sufficient opportunity for raters to observe the characteristics they are asked to judge.

For these reasons, behavior ratings by teachers or nursery school observers should be excellent data—and they appear to be. The unpublished Peterson-Cattell study reveals an average inter-judge reliability of .70 for single characteristics. Since a factor score

would be computed by a weighted pooling of perhaps five or six characteristics, one would expect inter-judge reliability for factor scores to be in the high .80's. This prediction is confirmed by Walker's (1962) study, in which two-thirds of the inter-judge reliabilities for nine cluster scores are in excess of .85. Walker also obtained ratings by different judges one year apart. They averaged a very respectable .56. It would appear unlikely that we shall find presently, or in the near future, behavior measures of much greater dependability.

Validity is something else again. Although the present study, together with other, similar studies, presents some evidence for construct validity, further evidence is necessary. One approach would be to regard behavior ratings as constituting the criterion domain, and to search for relationships with antecedents, such as birth-order, social class, etc. Another approach would be more in keeping with the history of scientific measurement: to regard ratings as rather good first approximations to behavior measurement, which will gradually give way to more objective instruments. In this connection, it should be noted that astronomy, generally regarded as the queen of the sciences, did not put its measurement system on a completely objective basis until well into the present century, but used for many years what would today be called "psychological scaling methods" for such basic data as star magnitudes. While all this "subjective" measurement was going on, the advances in astronomy were substantial.

REFERENCES

- Adler, A. *Social Interest*. New York: Putnam, 1939.
- Cattell, R. B. *Personality and Motivation Structure and Measurement*. Yonkers: World Book, 1957.
- Cattell, R. B. and Coan, R. W. "Child Personality Structure as Revealed in Teachers' Ratings." *Journal of Clinical Psychology*, XIII (1957), 315-327.
- Cattell, R. B. and Gruen, W. "The Personality Factor-Structure of 11-Year-Old Children in Terms of Behavior Rating Data." *Journal of Clinical Psychology*, IX (1953), 256-266.
- Digman, J. M. "Principal Dimensions of Child Personality as Inferred from Teachers' Judgments." *Child Development*, XXXIV (1963), 43-60. (a)
- Digman, J. M. "The Evidence for Complexity." Contribution to symposium, "Factor-Analytic Models of Child Personality: Simplicity vs. Complexity." Convention of the Society for Research in Child Development, Berkeley, 1963. (b)

- Dollard, J. and Miller, N. E. *Personality and Psychotherapy*. New York: McGraw-Hill, 1950.
- Eysenck, H. J. *The structure of Human Personality*. New York: Wiley, 1953.
- Guilford, J. P. *Psychometric Methods*. New York: McGraw-Hill, 1954.
- Peterson, D. R. "The Age Generality of Personality Factors Derived from Ratings." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XX (1960), 461-474.
- Peterson, D. R. and Cattell, R. B. "Personality Factors in Nursery School Children as Derived from Teachers' Ratings." Urbana, Ill.: The Laboratory of Personality Assessment and Group Behavior, University of Illinois, 1958 (mimeographed).
- Richards, T. W. and Simons, M. P. "The Fels Child Behavior Scales." *Genetic Psychology Monographs*, XXIV (1941), 259-309.
- Schaefer, E. S. "Converging Conceptual Models for Maternal Behavior and for Child Behavior." In J. C. Glidewell (Ed.), *Parental Attitudes and Child Behavior*. Springfield, Ill. Thomas, 1961.
- Stott, L. "Personality at Age Four." *Child Development*, XXX (1962), 287-311.
- Thurstone, L. L. *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947.
- Walker, R. N. "Body Build and Behavior in Young Children." *Monographs of the Society for Research in Child Development*, XXVII (1962), No. 3.



ACQUIESCENCE IN THE MMPI?¹

LEONARD G. RORER

Oregon Research Institute

AND

LEWIS R. GOLDBERG

University of Oregon and Oregon Research Institute

RESPONSES to personality inventory items have traditionally been thought of as "multiply determined," but only recently have the response determinants which are variously referred to as "sets," "biases," and "styles" been held to be more important than item content in accounting for such responses (e.g., Berg, 1959, 1961; Christie and Lindauer, 1963; Jackson and Messick, 1958; Loevinger, 1959; McGee, 1962c; Messick and Jackson, 1961). Recently, Rorer (1965) has proposed a conceptual distinction between "sets" and "styles." The term "set" is used to refer to the criteria according to which a respondent evaluates item content when selecting his answer. Sets have been designated by such terms as "dissimulation," "defensiveness," and "social desirability." The term "style" is used to refer to a way or manner of responding, such as a tendency to select some particular response option independently of the item content. Styles have been described by such terms as "yeasaying," "naysaying," and "extreme position response bias." Sets operate in relation to meaningful item content; styles operate in the absence

¹ This article is based on a portion of the first author's doctoral dissertation (Rorer, 1963). This study was supported in part by National Science Foundation Grant G-25123 to Oregon Research Institute under the direction of the second author. The analysis of much of the data was carried out through the facilities of Western Data Processing Center at the University of California at Los Angeles. The authors acknowledge with thanks the assistance of Dr. Paul E. Meehl, who made a number of helpful suggestions for improving an earlier draft of this article.

of such content. In the current literature "acquiescence" has been conceptualized as a generalized tendency to be agreeable (a set), but has been operationally defined in terms of a disproportionate tendency to select a certain response category (a style).

On the basis of an extensive review of the literature, Rorer (1965) concluded that, current opinion to the contrary notwithstanding there is no evidence that acquiescence responses style accounts for a significant proportion of the response variance in present personality inventories. Response styles could be established as an important response determinant either (a) by showing that two or more "content-independent" stylistic measures are significantly related, or (b) by showing that inconsistent responses are given to the same content presented in more than one form. Studies in which content-independent measures of style have been intercorrelated have found the measures to be unrelated—a result that holds for instruments which purportedly measure the same style as well as for instruments which purportedly measure different styles (e.g., Bass, 1956; Forehand, 1962; Gray and Crisp, 1961; Husek, 1961; McGee, 1962a, 1962b, 1962c; Siller and Chipman, 1962). These uniformly negative results have led to the hypothesis that response styles are "test specific." The establishment of response styles as an important response determinant in personality inventories rests, therefore, on the "reversed-content" design. In this design a respondent is presented with an item and with the reversal (logical contradictory) of that item. To the extent that correlations between responses to the original and the reversed items are high and negative (approaching the reliability of the test), individuals are responding consistently to the content of the items. To the extent that the correlations are high and positive, individuals are responding to a particular response category rather than to the content of the items. Most of the previous studies employing this design have utilized the California F scale. All have obtained intermediate correlations between the original and reversed forms of the test, and all have concluded that content and style are both important response determinants, though there is some difference of opinion as to which is the more important (e.g., Peabody, 1961). Rorer (1965) has criticized these studies on the grounds that the reversals employed were inadequate, and has concluded that they do not provide sufficient evidence to warrant the conclusion that re-

sponse styles are an important response determinant. However, neither do they offer evidence for rejecting that conclusion. Furthermore, there is in the literature today almost complete unanimity of opinion concerning the importance of response styles. It therefore seems that the weight of presumptive evidence in favor of the importance of response styles has become so strong in the minds of most investigators that it is incumbent upon one who would challenge that position to demonstrate that adequate item reversals can be written so that even test-specific response styles can be shown to be unimportant in determining responses to personality inventories.

Method

Procedure

An experimental group composed of 96 male and 125 female sophomores, juniors, and seniors from four psychology classes at the University of Minnesota was given the original Minnesota Multiphasic Personality Inventory (MMPI) and a reversed form of the MMPI two weeks apart. A control group composed of 95 male and 108 female students from an introductory psychology class at the University of Oregon was given the regular MMPI twice under similar conditions. The experimental and control groups will also be referred to as the reversal and reliability groups, respectively.

Instruments

In view of the previous research in this area, there were two tests which it seemed reasonable to use for this experiment—the California F scale and the MMPI. For both instruments, previous researchers had concluded that the proportion of the response variance attributable to styles was greater than that attributable to item content. Because it provided a larger, more diversified item pool, the MMPI was selected.

The development of the reversed items proved to be the most difficult part of the study. Conceptually the task was clear enough. Each of the original MMPI items had to be rewritten so that no matter in what way a respondent would answer the original item (and no matter what his reason for answering it in that way) he would be forced to answer the new item in the opposite way in order

to say the same thing. To put it another way, if a respondent were to be presented with the original and the reversed items simultaneously he would have to answer them in opposite directions if he were to avoid contradicting himself.

Items constructed in this way are reversals of the originals, not opposites. For example, in everyday speech the opposite of "black" is "white," but the reverse of "black" is "white and all shades of grey up to but not including black." The opposite of "loving" is "hating"; the reverse of "loving" is simply "an absence of loving," which may mean "hating," but also may simply mean indifference or no relationship at all. The opposite of "all" is "none"; the reversal is "any number excluding at least one." In other words, the reverse of a statement is that statement's logical contradictory, not its logical contrary or subcontrary. Two statements are contraries if they might both be false, but cannot both be true; two statements are subcontraries if they might both be true, but cannot both be false; two statements are contradictories if one must be true and the other false. A universal affirmative and a particular negative are contradictories, as are a universal negative and a particular affirmative, but a universal affirmative and a universal negative are not contradictories, and neither are a particular affirmative and a particular negative (see Copi, 1954, pp. 66-74).

If item reversal writing could be treated as the simple, logical task that it is conceptually, it would be a trivial exercise in symbolic logic. One could, for example, preface each item with "It is not the case that . . ." or "It is not true that. . . ." Unfortunately, prefixes such as this would also in most cases alter the item style, length and complexity simultaneously, and so are to be avoided if possible. Except for being reversed, the new items should be as much like the originals as possible. In almost all cases this means that the reversal should be as simple as possible. Unfortunately, the original items are in varying degrees vague, indefinite, ambiguous, and idiomatic, which means that in many cases there are no simple reversals.

The problems that are encountered in item reversal writing are best indicated by example. "I like mechanics magazines" cannot be reversed by "I do not like mechanics magazines," because the latter implies dislike. An individual who has no particular feelings about mechanics magazines one way or the other may consistently

reject both items. "I seldom worry about my health" is undoubtedly most often rejected by individuals who have a tendency to worry about their health, but it may also be rejected by some individuals who never worry about their health, on the grounds that the item implies at least some worry. A related problem is encountered with an item such as "At times I feel like smashing things." Some individuals feel that one or two incidents comprise sufficient grounds for endorsement; others feel that "at times" implies a recurring phenomenon which could be inferred only on the basis of a sizable number of such incidents. The item "I cry easily" cannot be reversed by "I do not cry easily," because to most respondents the latter item implies that it is difficult to make them cry. On the other hand, the confusion caused by reversing a negative item such as "I do not tire quickly" by "It is not true that I do not tire quickly" should be readily apparent even for this short item.

Though seven judges² provided innumerable suggestions concerning item interpretation and phrasing, the selection of the items to be included in the reversed form was ultimately based on the first author's notions concerning the relative importance of logical consistency as opposed to stylistic simplicity. Because the former was given more weight than the latter, the reversed items are probably appropriate only for college-level groups.³ For a more detailed account see Rorer (1963).

If the final results show that respondents given both item forms are as consistent as respondents given the same item twice, then there is no problem with this procedure. The experimentals have had to change their answer in order to be consistent, and they have been as consistent as the controls who were given the same item twice. The experimentals must be responding to the content. However, should the results show lower consistency for the experimentals, then the results might be attributable either to response bias or to inadequate reversals.

² The authors are indebted to Dr. and Mrs. Philip Gough, Dr. Starke R. Hathaway, Mrs. Denise Hawkins, Dr. and Mrs. James C. Kincannon, and Dr. Gail La Forge, all of whom spent much time evaluating preliminary item reversals. Whatever merit the final item reversals have rests heavily on the combined critical comments of these individuals.

³ A copy of the reversed MMPI may be obtained from the authors, without charge, or from the American Documentation Institute. Order Document No. 8454, remitting \$2.00 for 35 mm. microfilm, or \$3.75 for 6 x 8 photocopies.

Analysis

The results for males and females were analyzed separately. This split results in four groups which will be designated C-M (control males who took the original MMPI twice), C-F (control females who took the original MMPI twice), E-M (experimental males who took the original MMPI and the reversed MMPI), and E-F (experimental females who took the original MMPI and the reversed MMPI). For each of the four groups, the following statistics were calculated for each item on an IBM 7090 computer.

TT—The percentage of subjects responding "true" on both administrations of the item. (For the reversal groups, the subjects have actually marked "false" on the second administration. They are listed in this way because they have consistently responded to the item content and are to be compared with the reliability group which has consistently responded to the item content by marking "true" both times.)

TF—The percentage of subjects responding "true" on the first administration and "false" on the second administration of the item. (In this case, the reversal groups have actually marked "true" on both administrations of the test. These are the potential "acquiescence" responses.)

TB—The percentage of subjects responding "true" on the first administration and omitting a response on the second administration of the item. (In this case, no adjustment is made for the responses of the reversal groups.)

Qualifications concerning the reversed scoring of the reversal groups will not be continued in the following descriptions, but it is essential to remember that throughout the analyses the answers to the reversed form of the MMPI have also been reversed.

FT—The percentage of subjects responding "false" on the first administration and "true" on the second administration of the item.

FF—The percentage of subjects responding "false" on both administrations of the item.

FB—The percentage of subjects responding "false" on the first administration and omitting a response on the second administration of the item.

BT—The percentage of subjects omitting a response on the first administration and responding "true" on the second administration of the item.

BF—The percentage of subjects omitting a response on the first administration and responding "false" on the second administration of the item.

BB—The percentage of subjects omitting a response on both administrations of the item.

Blank-1—(BT + BF + BB) The percentage of subjects who failed to respond to the item on the first administration.

Blank-2—(TB + FB + BB) The percentage of subjects who failed to respond to the item on the second administration.

End %-1—(TT + TF + TB) The percentage of subjects responding "true" to the item on the first administration.

End %-2—(TT + FT + BT) The percentage of subjects responding "true" to the item on the second administration.

Shift—(End %-2—End %-1) The net shift (increase or decrease) between first and second administrations in the percentage of subjects endorsing the item.

Stable—(TT + FF + BB) The percentage of subjects who are consistent in their responses to the item.

Results

The detailed results of the analyses for each item over all subjects in each group are presented elsewhere (Goldberg and Rorer, 1963; Rorer, 1963). The mean of each of these statistics over all items is presented in Table 1. Because of rounding errors, there are .01 discrepancies in some of the results, but it should be noted that,

TABLE 1
Mean Values of Item Statistics over All Subjects and All Items

Index	Males		Females	
	C	E	C	E
TT	.35	.33	.35	.32
TF	.07	.08	.07	.07
FT	.06	.09	.06	.08
FF	.51	.50	.52	.52
Blank-1	.00	.01	.00	.00
Blank-2	.00	.00	.00	.00
End%-1	.42	.40	.42	.39
End%-2	.42	.42	.41	.41
Shift	-.00	.01	-.00	.01
Stable	.87	.83	.87	.84

Note—Values for TB, FB, BT, BF, and BB were all .00.

prior to rounding, the computer performed all calculations to eight-digit accuracy.

In general, the results for the various groups are similar. Inspection of the End %-1, End %-2, and Shift rows of the table shows that there was a 2 percent difference between the groups in over-all endorsement frequency on the first administration, and that this difference disappears on the second administration. This is due to the endorsement proportion for the controls decreasing slightly while the experimentals were increasing. In other words, the groups were more alike during the second testing when the experimentals were taking the reversals than during the first testing when both groups were taking the same items.

In order to assess the extent to which response styles rather than content account for the responses, it is necessary to compare the consistency of the experimentals with that of the controls. The values for Stable indicate that for both the male and female control groups, 87 percent of the responses were the same on both tests.

Because the reversed form of the MMPI was administered only once, no comparable test-retest stability values are available for it. However, for the 16 items which are repeated in the MMPI, intra-test stability values may be calculated for both forms. These values are shown in Table 2. It can be seen that the greater com-

TABLE 2
Within-Test Stability Values for 16 Repeated Items

Group	Males		Females	
	Admin. 1	Admin. 2	Admin. 1	Admin. 2
Controls	.94	.95	.95	.95
Experimentals	.96	.90	.95	.90

plexity of the reversed items has resulted in decreased response stability. If the relationship between inter- and intra-test stability for the reversed form is assumed proportional to that for the originals (a conservative assumption), then it would be predicted that 82 percent of the responses to the reversed form would be unchanged from one administration to another. If respondents are influenced entirely by item content, the stability of their responses when taking

both forms of the test would be expected to fall somewhere between 82 and 87 percent.

For the experimental groups in the present study, 83 percent of the responses of the males and 84 percent of the responses of the females were consistent; i.e., the subjects marked either TF or FT, indicating that they were responding to the item content rather than to the response category. Comparing the controls and the experimentals, there is a discrepancy of 4 percent for the males and 3 percent for the females, with greater response stability among the controls.

If the experimental subjects were acquiescing, then this discrepancy should be accounted for by a disproportionate percentage of "true" responses on the second administration by subjects who answered "true" on the first administration. Inspection of the table shows that for both the male and the female controls, 7 percent of the responses were changed from "true" on the first administration to "false" on the second administration. By comparison, for the reversal groups, 8 percent of the responses by the males and 7 percent of the responses by the females were "changed" from "true" to "true." In other words, using the most stringent possible criterion, double "trues" can account for only a .01 discrepancy in the males, and for none of the discrepancy in the females. What small discrepancy there is between the groups even on this criterion can almost all be accounted for by inconsistent double false responses by the reversal groups. When the greater instability of the reversed form is taken into account, even this "discrepancy" disappears. These results show quite clearly that acquiescence response style is of negligible importance in accounting for responses to the MMPI.

Figures 1 and 2 show scatter plots of the stability of each item for the experimentals and the controls. Figure 1 is for males and Figure 2 is for females. In Figure 1, a 45° line has been drawn to show the point at which items are equally stable for experimentals and controls. In Figure 2, the best fit regression line has been added. While this line may deviate from 45° because of real differences between the original and reversed items, it also may deviate from 45° because of the unreliability of the stability values. Since the obtained correlations (.64 for males and .61 for females) are as high as could be expected on the basis of existing estimates of the reliability of stability scores, it may be inferred that the stabi-

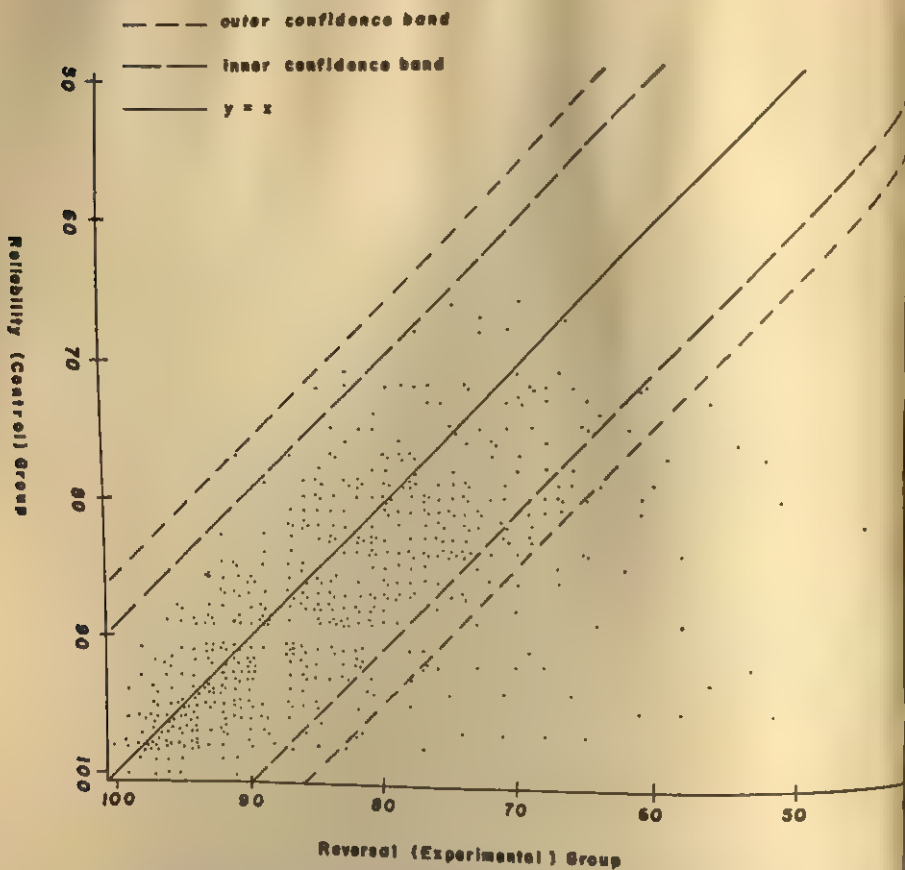


Fig. 1. Proportion of stable responses among males for each item. Each dot represents an item.

ties of the original and the reversed items are almost perfectly related.

Figures 1 and 2 show that items differed considerably in their stability, and that the reversals differed in their relative success. It is of interest to examine the least successful reversals in order to see if they possess any common characteristics. Variation about the 45° line was used as a selection criterion. Since there is no reason for the reversed items to be more stable than the originals administered twice, the range on the top side of the 45° line may be taken as an indication of the variation in item response stability that might occur by chance, and the same range was marked off on the bottom side of the 45° line. This will be termed the "outer confi-

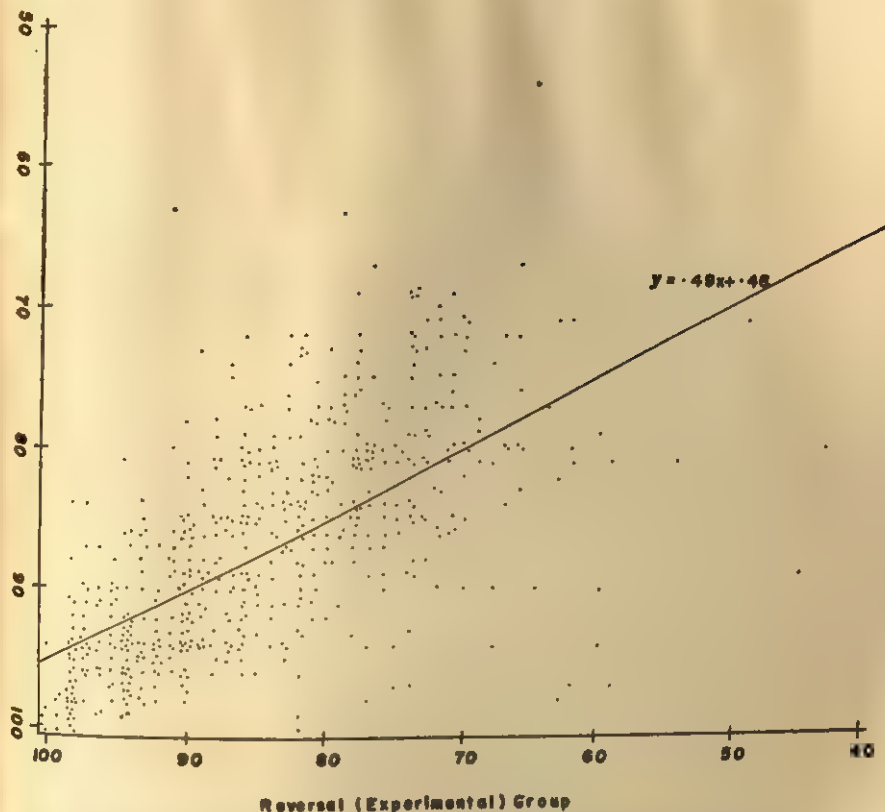


Fig. 2. Proportion of stable responses among females for each item. Each dot represents an item.

dence band." While the term "confidence band" is being used rather loosely, the empirically determined range seems as good as any of the possible alternatives. This value was 14 percent for the males and 16 percent for the females, but the smaller value was adopted for both groups. In addition to this outer confidence band, an inner band was drawn between 10 percent and 11 percent discrepancy. This procedure is illustrated in Figure 1.

In order to see if they shared some common characteristic, items were selected which (a) fell beyond the outer confidence band for both males and females, or (b) fell beyond the inner confidence band in both groups. The items selected in this manner are presented in Tables 3 and 4, where they are broken down into three groups: (a) items for which the instability was of the TT type;

(b) items for which the instability was of the FF type; (c) items for which the instability was about equally of the TT and FF type.

TABLE 3

Reversals for Which the Percentage of Stable Responses Was at Least 14 Percent Lower for the Reversal Sample Than for the Reliability Sample

True-True

- | | |
|--|---|
| 2. I have a good appetite. | I do not have a good appetite (or it is about average). |
| 17. My father was a good man. | My father was not a good man (or he was about average). |
| 83. Any man who is able and willing to work hard has a good chance of succeeding. | Not every man who is able and willing to work hard has a good chance of succeeding. |
| 127. I know who is responsible for most of my troubles. | I know of no one who is responsible for most of my troubles. |
| 175. I seldom or never have dizzy spells. | It is not true that I seldom or never have dizzy spells. |
| 460. I have used alcohol moderately (or not at all). | I have used alcohol immoderately. |
| 479. I do not mind meeting strangers. | It is not true that I do not mind meeting strangers. |
| 499. I must admit that I have at times been worried beyond reason over something that really did not matter. | I have never been worried beyond reason over something that really did not matter. |
| 513. I think Lincoln was greater than Washington. | I think Lincoln was no greater than Washington. |

Both True-True and False-False

- | | |
|---|---|
| 254. I like to be with a crowd who play jokes on one another. | I do not like to be with a crowd who play jokes on one another (or I am indifferent to it). |
| 441. I like tall women. | I do not like tall women (or I neither like nor dislike them). |

False-False

- | | |
|---|---|
| 16 (and 315). I am sure I get a raw deal from life. | I am not sure I get a raw deal from life. |
| 49. It would be better if almost all laws were thrown away. | It would be no better if almost all laws were thrown away. |
| 104. I don't seem to care what happens to me. | I do seem to care what happens to me. |
| 205. At times it has been impossible for me to keep from stealing or shoplifting something. | It has never been impossible for me to keep from stealing or shoplifting something. |
| 218. It does not bother me particularly to see animals suffer. | It is not true that it does not bother me particularly to see animals suffer. |
| 250. I don't blame anyone for trying to grab everything he can get in this world. | I do blame some people for trying to grab everything they can get in this world. |
| 265. It is safer to trust nobody. | It is no safer to trust nobody. |

284. I am sure I am being talked about.

287. I have very few fears compared to my friends.

375. When I am feeling very happy and active, someone who is blue or low will spoil it all.

387. The only miracles I know of are simply tricks that people play on one another.

461. I find it hard to set aside a task that I have undertaken, even for a short time.

I am not sure I am being talked about.

I do not feel that I have very few fears compared to my friends.

When I am feeling very happy and active, even someone who is blue or low cannot spoil it.

I know of miracles that are more than simply tricks that people play on one another.

It is not true that when I have undertaken a task, I find it hard to set it aside for a short time.

TABLE 4

Reversals for Which the Percentage of Stable Responses Was at Least 10 Percent Lower for the Reversal Sample Than for the Reliability Sample

True-True

64. I sometimes keep on at a thing until others lose their patience with me.

176. I do not have a great fear of snakes.

498. It is always a good thing to be frank.

563. I like adventure stories better than romantic stories.

I do not keep on at a thing until others lose their patience with me.

I can't honestly say that I do not have a great fear of snakes.

It is not always a good thing to be frank.

I like adventure stories no better than romantic stories.

Both True-True and False-False

58. Everything is turning out just like the prophets of the Bible said it would.

399. I am not easily angered.

523. I practically never blush.

Not everything is turning out just like the prophets of the Bible said it would.

It is not true that I am not easily angered.

It is not true that I practically never blush.

False-False

10. There seems to be a lump in my throat much of the time.

14. I have diarrhea once a month or more.

39. At times I feel like smashing things.

100. I have met problems so full of possibilities that I have been unable to make up my mind about them.

198. I daydream very little.

328. I find it hard to keep my mind on a task or job.

It is not true that there seems to be a lump in my throat much of the time.

I have diarrhea less than once a month.

I never feel like smashing things.

I have never met problems so full of possibilities that I have been unable to make up my mind about them.

I daydream more than a very little. Maybe some people find it hard to keep their mind on a task or job, but not me.

The items in Tables 3 and 4 are heterogeneous with respect to scale membership, content, and form of reversal. Inspection indicates only one common characteristic: TT items tend to have high initial endorsement frequencies, FF items tend to have low initial endorsement frequencies, and mixed items tend to have medium endorsement frequencies. Inspection of the items indicates that their instability most likely stems from either misunderstanding on the part of the respondents (e.g., 175, 460, 16, and 218) or inadequacies inherent in the reversals (e.g., 2, 17, 83, 205, and 499).

Discussion

The response stabilities obtained for the experimentals and controls in this study may be compared with values previously reported. Schofield (1948) found that a group of normals were consistent in 86 percent of their responses to 495 MMPI items on two different occasions. Schofield also reported stability percentages of 78 percent for psychiatric out-patients, 82 percent for hospitalized neurotics, and 69 percent for psychotics. Neprash (1936) found that after two weeks' time, 85.9 percent of his subjects' responses to the Thurstone Personality Schedule were unchanged. Benton and Stone (1937) found that for Landis and Zubin's Personality Inquiry Form, which also has a yes-?-no format, 81 percent of their subjects' responses were unchanged after five days, 80 percent after eight days, and 81 percent after 21 days. In other words, the failure to find acquiescence in this study cannot be attributed to an inconsistent or poorly motivated control group. The controls in this study were more consistent in their responses than any other group to be found in the literature.

When this fact is considered in conjunction with the undoubtedly lower test-retest stability of the reversed form, and the inadequacies of some of the reversals, it seems remarkable that the consistency of the experimentals could be even close to that of the controls, much less identical to it (as was the case for the TT or "acquiescent" category of responses for the females).

Perhaps the point is best made in reverse. If one wishes to interpret these results in terms of response styles, and if he is willing to give *no* weight to: (a) the possibility of some discrepancies due to sampling error, (b) the extreme consistency of this control group in relation to all other groups reported in the literature, (c) the un-

doubtedly lower test-retest stability of the reversed MMPI, and (d) the possibility that some of the reversals are not perfect, then it could be said that 1 percent of all responses made by males are "acquiescent," and 3 percent of all responses by males and 2 percent of all responses by females are "critical." Absurd as this line of reasoning is, the amount of response style variance which it manages to salvage is trivially small.

It has become fashionable to interpret the first two orthogonal MMPI factors in terms of acquiescence and social desirability, and to give response style names to scales constructed on the basis of certain item characteristics such as endorsement percentage or direction of keying. This naming and interpretation is necessarily equivocal. Since the scales are composed of verbal items, the respondent may be giving factual answers, no matter how many of them are true or false.

The fact that an individual gives a preponderance of "true" responses to the items on a personality inventory, or even to any subset of items from an inventory, provides no basis on which to conclude that he "acquiesced." For personality, attitude, and interest inventories, unless the items are stated in more than one way, the content and the keying are inevitably confounded, no matter how much statistical legerdemain is performed upon the results. Previous studies concluding that acquiescence response style is an important variable in determining MMPI responses have all done so on the basis of such content-confounded measures.

Such conclusions are no longer tenable. Chapman and Campbell (1959) have previously shown that adequate reversals could be written for the MMPI at scale, and Block (1965) has shown that MMPI scales balanced for "true" and "false" keyings yield the same factor structure as the present unbalanced scales. Factors extracted using balanced scales obviously cannot be interpreted in terms of acquiescence or criticalness response styles. The results of these investigators, taken in conjunction with the results of this experiment, indicate quite clearly that acquiescence and criticalness response styles are of negligible importance in determining MMPI responses for the groups that have been studied. It is conceivable that this would not be the case among other groups. Scales might be discriminative because certain groups, such as college students, lack acquiescence and criticalness set, whereas various

clinical groups have one or the other. However, in a study aimed at precisely this question, Jackson and Messick (1962) concluded that the "massive" effects of response styles were almost identical for college students, prison inmates, and hospitalized neurotics. Since their results indicate the same effects for all three groups, this study may be similarly generalized and serves to call all of them into question.

Summary

Reversals (logical contradictories) were written for all MMPI items. Responses of an experimental group given the original MMPI and the reversed MMPI were compared with those of a control group given the original MMPI twice. The results indicate that "acquiescence response style" can be of no more than trivial importance in determining responses to the MMPI.

REFERENCES

- Bass, B. M. "Development and Evaluation of a Scale for Measuring Social Acquiescence." *Journal of Abnormal and Social Psychology*, LIII (1956), 296-299.
- Benton, A. L. and Stone, I. R. "Consistency of Response to Personality Inventory Items as a Function of Interval between Test and Retest." *Journal of Social Psychology*, VIII (1937), 143-146.
- Berg, I. A. "The Unimportance of Test Item Content." In B. M. Bass and I. A. Berg (Editors), *Objective Approaches to Personality Assessment*. New York: Van Nostrand, 1959.
- Berg, I. A. "Measuring Deviant Behavior by Means of Deviant Response Sets." In I. A. Berg and B. M. Bass (Editors), *Conformity and Deviation*. New York: Harper and Brothers, 1961.
- Block, J. *The Challenge of Response Sets*. New York: Appleton-Century-Crofts, 1965.
- Chapman, L. J. and Campbell, D. T. "Absence of Acquiescence Response Set in the Taylor Manifest Anxiety Scale." *Journal of Consulting Psychology*, XXIII (1959), 465-466.
- Copi, I. M. *Symbolic Logic*. New York: The Macmillan Company, 1954.
- Christie, R. and Lindauer, F. "Personality Structure." *Annual Review of Psychology*, XIV (1963), 201-238.
- Forehand, G. A. "Relationships among Response Sets and Cognitive Behaviors." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 287-302.
- Goldberg, L. R. and Rorer, L. G. "Test-retest Item Statistics for Original and Reversed MMPI Items." *Oregon Research Institute Research Monograph*, III, No. 1 (1963).

- Gray, C. W. and Crisp, H. E. "The Credibility of Pure Response Set." Paper read at the annual meeting of the Southeastern Psychological Association, Gatlinburg, Tennessee, 1961.
- Husek, T. R. "Acquiescence as a Response Set and as a Personality Characteristic." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 295-308.
- Jackson, D. N. and Messick, S. "Content and Style in Personality Assessment." *Psychological Bulletin*, LV (1958), 243-252.
- Jackson, D. N. and Messick, S. "Response Styles on the MMPI: Comparison of Clinical and Normal Samples." *Journal of Abnormal and Social Psychology*, LXV (1962), 285-299.
- Loevinger, Jane. "Theory and Techniques of Assessment." *Annual Review of Psychology*, X (1959), 287-316.
- McGee, R. K. "The Relationship between Response Style and Personality Variables: I. The Measurement of Response Acquiescence." *Journal of Abnormal and Social Psychology*, LXIV (1962), 229-233. (a)
- McGee, R. K. "The Relationship between Response Style and Personality Variables: II. The Prediction of Independent Conformity Behavior." *Journal of Abnormal and Social Psychology*, LXV (1962), 347-351. (b)
- McGee, R. K. "Response Style as a Personality Variable: By What Criterion?" *Psychological Bulletin*, LIX (1962), 284-295. (c)
- Messick, S. and Jackson, D. N. "Acquiescence and the Factorial Interpretation of the MMPI." *Psychological Bulletin*, LVIII (1961), 299-305.
- Neprash, J. A. "Reliability of Questions in the Thurstone Personality Schedule." *Journal of Social Psychology*, VII (1936), 239-244.
- Peabody, D. "Attitude Content and Agreement Set in Scales of Authoritarianism, Dogmatism, Antisemitism, and Economic Conservatism." *Journal of Abnormal and Social Psychology*, LXIII (1961), 1-11.
- Rorer, L. G. "The Function of Item Content in MMPI Responses." Unpublished doctoral dissertation, University of Minnesota, Minneapolis, 1963.
- Rorer, L. G. "The Great Response Style Myth." *Psychological Review*, LXIII (1965), 129-156.
- Schofield, W. S. "MMPI Changes with Certain Therapies." Unpublished doctoral dissertation, University of Minnesota, Minneapolis, 1948.
- Siller, J. and Chipman, A. "Response Set Paralysis: Implications for Measurement and Control." *American Psychologist*, XVII (1962), 391. (Abstract)

ACCEPTANCE OF *Sc* SCALE STATEMENTS BY VISUAL ART STUDENTS¹

IRWIN J. GOLDMAN

Columbia University²

HIGHER scores on the MMPI have been linked to interest in the visual arts (Redlo, 1951; Spiaggia, 1950; Sternberg, 1953). Two possible explanations of this connection are (a) persons concerned with the visual arts tend to possess characteristics also present in psychopathology, these characteristics denoted by MMPI item content, and (b) persons concerned with the visual arts tend to be more willing to affirm the kinds of statements which characterize the MMPI, i.e. socially undesirable, pathology-connoting statements. In MMPI studies of aesthetically inclined individuals of diverse types (Barron, 1953; Chyatte, 1949; Keston, 1956; Redlo, 1951; Spiaggia, 1950; Sternberg, 1953; Taft, 1961) the most consistently discriminating clinical scale other than *Mf* (Interest) has been the *Sc* (Schizophrenia) scale.³ This scale has been found greatly affected by social desirability tendency (Fordyce, 1956).

A means of separating the two bases of MMPI item affirmation lies in the use of the forced choice format (Edwards, 1957). Ac-

¹ This paper is based on a doctoral dissertation submitted to the faculty of Columbia University in February, 1962. Grateful acknowledgment is made to the sponsor of the study, Richard Christie, and to members of the dissertation committee, Otto Klineberg and Marshall Segall. The writer is indebted to the officers, faculty, and students of schools participating in this study for their support and cooperation.

² Now with the New York State Division for Youth.

³ The *Mf* scale is not generally considered a measure of psychopathology. The "artistic types" were actors, actresses, visual art students, music students, and high-scorers on these measures: Barron-Welsh Art scale, Keston Music Preference test, Kuder Preference Record Artistic, Music, and Literary scales, Allport-Vernon Study of Values Aesthetic scale.

cordingly, for the purpose of further exploring the propensity of persons concerned with the visual arts in endorsing MMPI items, a forced choice questionnaire with Sc and non-Sc statements was constructed (Goldman, 1962).

Chyatte (1949) had reported an MMPI profile for actors almost identical with that reported for visual art students (Spiaggia, 1950). The protocols of these actors were available for reanalysis with the questions of this study in mind (Chyatte and Goldman, 1961). The actors showed no tendency to reveal socially unfavorable, pathology-connoting traits, but rather a tendency to conceal such characteristics. On the basis of this and the findings of the previously cited MMPI studies it was expected that persons concerned with the visual arts would not exhibit an unusual willingness to admit socially unfavorable characteristics. It was, however, expected that these persons would tend to affirm the content of Sc items when social desirability was controlled.

This study represented an attempt to (a) further explore and test the generality of the connection between interest in the visual arts and MMPI item affirmation, and, if reoccurring in this study, (b) specify its basis in terms of social desirability tendency and/or item content acceptance.

Procedure

Subjects. The art group consisted of students attending a school of visual art in New York City, a highly respected, degree-granting institution accredited by the Middle State Association of Colleges and Secondary Schools. The vast majority of its graduates entered art professions according to school placement surveys. The comparison group consisted of students at four liberal arts colleges in the New York City environs (Columbia, Barnard, Hofstra, Newark State) whose major field of study was not an art field. Art and nonart subjects were matched in age, sex, year at college, and father's formal education (an indicator of socioeconomic status).

Measures. The following six measures were used. (1) *Y-N Sc.* This comprised 19 statements from the Sc scale presented to respondents in a "Yes-No" format. The "Yes-No" format used for this and other measures refers to the presentation of statements for agreement-disagreement along a five-point scale (from "definitely yes" to "definitely no").

(2) *F-C Sc*. The same statements were presented in a forced choice format similar to that used by Heineman (1953). Each item consisted of three statements, two matched in social desirability value and the third of a contrasting value. Only the matched statements, one an *Sc* statement and the other a non-*Sc* statement, were compared for scoring. Matchings were based on ratings of 82 undergraduate and 29 graduate students with statements matched for both groups of raters. Subsequent to the study Messick and Jackson (1961) presented social desirability scale values for MMPI items using the scaling method of successive intervals. Using their values, the matched statements of *F-C Sc* are still closely matched with the mean difference in desirability for all items .21 scale units of a nine-point scale.⁴

The 19 *Sc* statements were selected from the full 78-item *Sc* scale by their greater absolute frequency of endorsement by actors in a previous study (Chyatte, 1949). The items were thereby among the most frequently endorsed *Sc* items in general, accounting for over 40 percent of all *Sc* scale endorsements by college students in one study (Appendix E in Dahlstrom and Welsh, 1960), and may also be of a kind more likely to be endorsed by art-oriented persons.⁵ The matched non-*Sc* statements (with two exceptions) came from the other clinical scales.

(3) *Discrepancy*. The art and nonart groups were combined and scores on *Y-N Sc* and on *F-C Sc* were rank ordered. The difference in rank order for an individual between *Y-N Sc* and *F-C Sc* was called *Discrepancy* and was used as a measure of social desirability tendency in reference to the *Sc* statements. High *Discrepancy* scores (indicating relatively greater affirmation of *Sc* statements in the forced choice format than in the Yes-No format) were interpreted as signifying concealment tendency.

(4) *M-C*. Twenty-five statements from the Marlowe-Crowne So-

⁴ A detailed description of this measure is to be found in Goldman (1962). Of the initial 19 items one item was dropped to maintain internal consistency. An example of an item:

A. Many of my dreams concern sex matters.

B. I have been disappointed in love.

C. I like to be independent in deciding what to do.

⁵ This was considered advantageous in assessing the influence of social desirability tendency on *Sc* endorsements by art-oriented persons. The MMPI items were also modified in form when placed in the Yes-No format, changed from a first person declarative to a second person interrogative sentence.

cial Desirability scale were presented in the Yes-No format. These statements are scored for socially desirable but improbable self-descriptions of a nonpathological nature (Crowne and Marlowe, 1960; Marlowe, 1961; Marlowe and Crowne, 1961). They were included to investigate social desirability tendencies in regard to nonpathological statements.

(5) *Non-Sc*. The 19 matched non-Sc statements of *F-C Sc* were presented in the Yes-No format as an aid in examining the source of any differences on *F-C Sc*.

(6) *Ratings*. The Sc statements were presented for social desirability ratings along a five-point scale in order to evaluate the effect on *Y-N Sc* and *F-C Sc* scores of group differences in standards of desirability with reference to the Sc statements.

These measures were contained in a single questionnaire containing several others not relevant to this study.

Administration. The questionnaire was given out in a usual class period and filled out by all students there. (The art students were in required Foundation of Art classes; the nonart students were in introductory psychology or sociology classes.) The study was presented as an investigation of the personality traits of college students. Subjects were promised further information regarding the study to enlist their cooperation. They were assured that individual replies would be considered confidential, and they were given the option of remaining anonymous if they wished. Generally, subjects appeared cooperative and interested. The administration took an hour.

Matching Subjects and Analysis. Subjects were matched in age, sex, year at college, and father's formal education. From a pool of 44 visual art and 115 nonart students each visual art student was

TABLE 1
Age and Sex Distributions of Art and Matched Groups

Age	Art Ss	Matches
17 & 17.5	2	2
18 & 18.5	27	27
19 & 19.5	8	9
20	5	3
21-25	2	3
Sex		
Male	25	25
Female	19	19

TABLE 2

Distributions for Year at College and Father's Education

Year at College	Art <i>Sc</i>	Matches
First	40	35
Second	0	5
Third	1	3
Fourth	1	1
Unknown	2	0
Father's Education		
Under 7 grades	2	2
7-9 grades	3	5
10-11 grades	4	5
12 grades	7	10
Some college	10	10
Completed college	11	10
Graduate training	4	2
Unknown	3	0

closely matched with a nonart student. The groups were composed mainly of students in their first year of college (see Tables 1 and 2). The difference in score between matched students was the unit of analysis. Since distributions did not appear normal and measures were not considered at interval level, it was analyzed by means of nonparametric tests, i.e. the Sign test and the Wilcoxon matched-pairs signed-ranks test (Siegel, 1956). Two-tailed tests were used.

Results

Using the protocols of 84 liberal arts students and an odd-even split half method, reliability coefficients were computed for *Y-N Sc*, *Non-Sc*, and *M-C*; these were .79, .67, and .75, respectively. For the *F-C Sc* items the Kuder-Richardson formula 20 was used with the same protocols; the reliability coefficient was .33. The lowered figure for *F-C Sc*, compared to *Y-N Sc*, is probably due to the reduction of social desirability influence and the elimination of acquiescent response set. Evidence that the forced choice format was effective in minimizing social desirability tendency influence is presented in Goldman (1964). The 84 liberal arts students consisted of the 44 matched students of this study and 40 others from the same classes involved in a parallel study of music students.

The major findings, as presented in Table 3, were as follows. (1) The art group did not score higher on *Y-N Sc*. (2) The art group did not score higher on *Non-Sc*. (3) The art group did score

TABLE 3

Number of Pairs in which Art Ss Scored Higher and Lower than Matches

	Y-N Sc	F-C Sc	Discrepancy	M-C	Rating	Non-Sc
Higher	20	25	29	28	17	18
Lower	21	14	14	16	24	23
Sign test z-score	.00	1.60	2.14	1.66	.94	.62

Note—A z-score of 1.96 would be significant at the .05 level (two-tailed test). The Wilcoxon matched-pairs signed-ranks z-score for F-C Sc was 2.09.

significantly higher ($p < .05$) on F-C Sc.⁶ (4) The art group did score significantly higher ($p < .05$) on Discrepancy. (5) The art group scored noticeably higher, but not significantly, on M-C. (6) There were no significant differences in social desirability ratings of Sc statements.

Discussion

The visual art group of this study did not tend to affirm Sc statements in the Yes-No format, indicating limits to the generality of the connection of previous studies between interest in the visual arts and Sc item endorsement. Nevertheless, the visual art group did tend to affirm Sc statements in the forced choice format considerably more frequently than matched counterparts despite the apparently low reliability of the instrument. Since the visual art students did not tend to score higher on Y-N Sc, their higher scores on F-C Sc cannot be attributed to social desirability tendency. One may conclude that the visual art students tended to accept the self-descriptive content of the Sc items, and the initial expectation that this would occur is confirmed.

There were differences in social desirability tendency between the compared groups. Similar to the actors of Chyatte and Goldman (1961) the visual art students tended to deny (not reveal) pathology-connoting statements, as indicated by their significantly higher Discrepancy scores. The differences on M-C, though not statistically significant, were in the same direction as Discrepancy. The second of the initial expectations was therefore also confirmed.

However, the discrepancy between the findings of this study for

⁶ Differences were not significant with the Sign test but were with the more powerful Wilcoxon matched-pairs signed-ranks test. This was the only measure that the Wilcoxon z-score was substantially different from the Sign test z-score.

Y-N Sc and the previous findings must still be explained. In view of the *F-C Sc* scores it appears that social desirability tendency was so great for this particular visual art group that *Y-N Sc* differences did not appear. The subjects of this study differed from those of the previous ones in many ways, among the most important may be (1) they were younger, and relatively new to college life, (2) they were not volunteers, but a "captive audience" and (3) they included many students oriented to commercial art fields. Conceivably, all three variables might be associated with a heightened social desirability tendency. It was not possible from the data collected to evaluate the effects of age or volunteering, but it was possible to evaluate the effect of commercial art orientation.

Included in the questionnaire were biographical questions, two of which were "What is your major field of study?" and "What are your future professional plans?". Students were categorized as "commercially oriented" (CO) if they stated that their major field of study was commercial art and/or if they stated they planned to enter fields of commercial art, illustration, advertising, or industrial design. All other art students were classified as "non-commercially oriented" (NCO). The scores of these subgroups were then compared with reference to whether they were higher, lower, or equal to their matched nonart subjects. The results for the principal measures are given in Table 4.

TABLE 4

Comparison of Commercially Oriented and Noncommercially Oriented Visual Art Students

	<i>Y-N Sc</i>			<i>F-C Sc</i>			<i>M-C</i>			Discrepancy		
	Hi	Lo	Tie	Hi	Lo	Tie	Hi	Lo	Tie	Hi	Lo	Tie
CO	9	16	2	15	8	4	20	7	0	20	7	0
NCO	11	5	1	10	6	1	8	9	0	9	7	1

Note.—Comparison is with reference to nonart matched students. *Hi* signifies score higher than match; *Lo* signifies score lower than match.

The table shows that the CO art students did respond differently from NCO students on *Y-N Sc* ($\chi^2 = 4.19$, $p < .05$, two-tailed test and excluding tie scores). Sixty-five percent of NCO subjects scored higher than their matches on *Y-N Sc* as opposed to 33 percent of CO subjects. In their higher scores on *Y-N Sc* the NCO art students were similar to the visual art subjects of the previous

studies. Further, Table 4 shows that it was the CO student who tended to exhibit concealment tendencies; the CO students were significantly different ($p < .03$) from their matches on both measures of social desirability (M-C and Discrepancy). The NCO students were not characterized by social desirability tendency in either direction.

Despite differences between CO and NCO subjects on Y-N Sc there were no marked differences between the two subgroups on F-C Sc. Table 5 presents the Sc statements in the forced choice items in the order of degree of differences in endorsement between the visual art and matched group (as determined by chi-square). Fortunately for interpretation, the content of the Sc statements appear fairly homogeneous. In general the content of Sc statements which the visual art group tended to accept appear to refer to problems in the expression of impulse, an interpretation which

TABLE 5

Sc Statements in Forced Choice Items in Order of Frequency of Endorsement by Visual Art Ss Compared to Matched Ss

-
1. I have strange and peculiar thoughts. (3.03)*
 2. My sex life is satisfactory.^b (2.07)
 3. I am worried about sex matters. (1.86)
 4. I have had periods of days, weeks, or months when I couldn't take care of things because I couldn't "get going." (1.06)
 5. I have difficulty in starting to do things. (1.02)
 6. Once in a while I feel hate toward members of my family whom I usually love. (.96)
 7. I have been afraid of things or people that I knew could not hurt me. (.96)
 8. I dream frequently about things that are best kept to myself. (.71)
 9. I have periods of such great restlessness that I cannot sit long in a chair. (.69)
 10. Many of my dreams concern sex matters. (.46)
 11. At times I have a strong urge to do something harmful or shocking. (.29)
 12. During one period when I was a youngster, I engaged in petty thievery. (.15)
 13. Once in a while I think of things too bad to talk about. (.00)
 14. I wish I were not bothered by thoughts about sex. (.06)
 15. I refuse to play some games because I am not good at them. (.71)
 16. I love (or loved) my father.^b (1.47)
 17. My daily life is full of things that keep me interested.^b (1.71)
 18. I find it hard to keep my mind on a task or job. (1.86)
-

*Chi-square value, corrected for continuity, occurring under the null hypothesis.

^bScored for denial in the MMPI.

Note. Statements 1-12 were within items more frequently endorsed by visual art subjects; statements 14-18 were within items more frequently endorsed by matched subjects.

would accord with Myden's (1959) inferences based on projective tests that artists are closer to their impulse life than nonartists.

Summary

Previous studies suggested an association between interest in the visual arts and heightened MMPI score. To examine the generality and basis of this association visual art students were compared to matched nonart students on six measures. The visual art students tended to affirm *Sc* statements when forced to choose between two socially undesirable, pathology-connoting self-descriptions. They did not tend to affirm *Sc* statements in the Yes-No format. It was concluded that the visual art group tended to accept the self-descriptive content of the *Sc* statements, apart from social desirability tendency.

REFERENCES

- Barron, F. "Complexity-Simplicity as a Personality Dimension." *Journal of Abnormal and Social Psychology*, XLVIII (1953), 163-172.
- Chyatte, C. "Personality Traits of Professional Actors." *Occupations*, XXVII (1949), 285-288.
- Chyatte, C. and Goldman, I. J. "The Willingness of Actors to admit to Socially Undesirable Behavior on the MMPI." *Journal of Clinical Psychology*, XVII (1961), 44.
- Crowne, D. P. and Marlowe, D. "A New Measure of Social Desirability Independent of Psychopathology." *Journal of Consulting Psychology*, XXIV (1960), 349-354.
- Dahlstrom, W. G. and Welsh, G. S. *An MMPI Handbook*. Minneapolis: University of Minnesota Press, 1960.
- Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research*. New York: Dryden, 1957.
- Fordyce, W. E. "Social Desirability in the MMPI." *Journal of Consulting Psychology*, XX (1956), 171-175.
- Goldman, I. J. "The Willingness of Music and Visual Art Students to Admit to Socially Undesirable and Psychopathological Characteristics." Unpublished doctoral dissertation, Columbia University, 1962.
- Goldman, I. J. "Effectiveness of the Forced Choice Method in Minimizing Social Desirability Influence." *Journal of Consulting Psychology*, XXVIII (1964), 289.
- Heineman, C. E. "A Forced Choice Form of the Taylor Anxiety Scale." *Journal of Consulting Psychology*, XVII (1953), 447-454.
- Keston, M. "An Experimental Investigation of the Relationship between the Factors of the Minnesota Multiphasic Personality In-

- ventory and Musical Sophistication." *American Psychologist*, XI (1956), 434. (Abstract)
- Marlowe, D. "Need for Social Approval and the Operant Conditioning of Meaningful Verbal Behavior." *Journal of Consulting Psychology*, XXVI (1961), 79-83.
- Marlowe, D. and Crowne, D. P. "Social Desirability and Response to Perceived Situational Demands." *Journal of Consulting Psychology*, XXV (1961), 109-115.
- Messick, S. and Jackson, D. N. "Desirability Scale Values and Dispersions for MMPI Items." *Psychological Reports*, VIII (1961), 409-414.
- Myden, W. "Interpretation and Evaluation of Certain Personality Characteristics Involved in Creative Production." *Perceptual and Motor Skills*, IX (1959), 139-158.
- Redlo, M. "MMPI Personality Patterns for Several Academic Major Groups." Unpublished master's thesis, University of New Mexico, 1951.
- Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956.
- Spiaggia, M. "An Investigation of the Personality Traits of Art Students." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, X (1950), 284-293.
- Sternberg, C. "The Relation of Interests, Values and Personality to the Major Field of Study in College." Unpublished doctoral dissertation, New York University, 1953.
- Taft, R. "A Psychological Assessment of Professional Actors and Related Professions." *Genetic Psychology Monographs*, LXIV (1961), 309-383.

THE VALIDITY OF THE EDWARDS PERSONAL PREFERENCE SCHEDULE (EPPS) EMPLOYING PROJECTIVE AND BEHAVIORAL CRITERIA¹

DANIEL V. CAPUTO,² JON M. PLAPP
Washington University School of Medicine

CONSTANCE HANF
University of Oregon Medical School
AND ANNE SMITH ANZEL
University of California, Los Angeles

In the present study, two approaches were taken in evaluating the validity of the EPPS: (a) the test's ability to discriminate between groups, and (b) the extent to which there was agreement or correlation between EPPS needs and measures of these needs on other instruments, for the same Ss.

Investigations of the concurrent validity of the EPPS have produced conflicting results. Edwards (1959) presented evidence for EPPS validity based on agreement between self-ratings and test scores and on correlations between EPPS and other personality measures. Negative findings were obtained by Graine (1957), Allen (1957), Hartley and Allen (1962), and Marlowe (1959), while the findings of Dunette, Kirchner, and De Gidio (1958), Heilbrun (1958), Zuckerman and Grosz (1958), Sheldon, Coale, and Copple (1959), Krug and Moyer (1961), and Zuckerman, Levitt, and Lubin (1961) supported EPPS validity of some or all of the scales.

¹ This paper is a partial report of a research project "Role Differentials and Nursing Ideology," Drs. John A. Stern and Albert F. Wessen, Co-Principal Investigators; the research has been supported by National Institute of Health Grant NU-00050. A portion of the computation was done with support provided by the Washington University Computer Center under National Science Foundation Grant G-22296.

² Now at Queens College of the City University of New York.

In general, predictive validity studies employing the EPPS achievement scale have obtained positive results (Bendig, 1958; Gebhart and Hoyt, 1958; Krug, 1959; Mogar, 1962; Worell, 1960), although the findings of Goodstein and Heilbrun (1962) and of Lang, Sferra, and Seymour (1962) were equivocal.

Agreement between the constructs of dependency, rebelliousness, and resistance to change and certain EPPS scales has been reported by Zuckerman and Grosz (1958), Zuckerman (1958), Bernardin and Jessor (1957), and Izard (1960).

Certain methodological issues relevant to the investigation of EPPS validity have been studied. Edwards' method of controlling for the social desirability factor has been found to yield stable results across different groups (Klett, 1957a, 1957b; Klett and Tamkin, 1957; Klett and Yaukey, 1959). However, Heilbrun and Goodstein (1959) stated that there may be some independent contribution of a personal desirability set beyond that attributable to a social desirability set. A later study by the same authors (Heilbrun and Goodstein, 1961) demonstrated empirically that prediction was not enhanced by reducing the effect of social desirability, and in fact it was felt that this factor may be a source of test validity.

Method

The major strategy of the present investigation was to form two clearly differentiated groups on the basis of Ss' scores on certain EPPS needs, and then to determine whether other methods of assessing these same needs differentiated the groups in the same manner or related to the EPPS in the predicted way.

Subjects consisted of 30 girls selected from the 1962-3 freshman nursing student class of a large general hospital.

The EPPS was the instrument used for the initial selection of groups. The other instruments used in the assessment of Ss were the Role Projective Test (RPT), the Interaction Situation (IS), and the Role Behavior Test (RBT).

The Role Projective Test is an 11-card TAT-like test designed by Wessen and first employed by Barber and Wessen (in press). Each card depicts a nursing situation. In the present investigation four cards depicting a nursing student in the following four situations were used: attending a male patient, with a medical intern,

with a nursing supervisor, about to be summoned by a patient.

The Interaction Situation consisted of a pair of Ss discussing certain of their responses to a previously administered 70-item questionnaire assessing nursing student attitudes. The girls were physically separated and communication between them established by means of an intercom system. Discussions were tape recorded.

The Role Behavior Test consisted of role playing situations which required Ss to respond to a simulated sick male patient, medical intern, nursing supervisor, and nursing student peer. Appropriate props were employed to enhance the reality of the situations.

Two experimental rooms, with a one-way mirror between them, were used. The critical incidents occurred in one room while the recording and observing activities occurred in the second. The verbal interactions were recorded on one track of a four-track tape recorder with verbal comments of observer recorded on a second track.

Procedure

The EPPS was administered in the standard manner to the entire class approximately two months after admission to nursing school.

On the basis of their scores on the EPPS two groups, each consisting of 15 Ss, were formed in the following manner.

From the EPPS scores of the total group of 79 nursing students those five needs with highest group means and those five needs with the lowest group means were determined. The resulting five highest needs were nurturance, intraception, abasement, change, and affiliation; the five lowest needs were dominance, achievement, autonomy, order, and aggression.

Those Ss whose scores on any of the *high* needs were at least .75 SD *higher* than the mean of the total group and those Ss whose scores on any of the *low* needs were at least .75 SD *lower* than the mean of the total group were selected. They were then ranked, those Ss who deviated by the required amount of .75 SD on the greatest number of needs (both high and low) ranking highest, and those Ss who deviated by the required amount on only one need ranking lowest. The 15 highest ranking Ss were then selected (Gp I Hi Hi—Lo Lo).

A second group whose scores on any of the *high* needs were at least .75 SD *lower* than the means of the total group and those *Ss* whose scores on any of the *low* needs were at least .75 SD *higher* than the means of the total group were also chosen from the original group of 79 nursing students. They were then ranked, with those *Ss* who deviated by the required amount of .75 SD on the greatest number of needs (both high and low) ranking highest, and those *Ss* who deviated by the required amount on only one need ranking lowest. The 15 highest-ranking *Ss* were then selected (Gp II Hi Lo—Lo Hi).

These two groups were therefore chosen in such a way that the differences between them, in terms of the 10 need scores, were maximized. The two groups were then tested on the criterion instruments: the RPT, the IS, and the RBT.

In order to determine whether the differences between G-I and G-II means on each of the 10 needs were significant, *t*-tests were performed. The two groups were significantly different on each need (at or beyond the .05 level).

The Role Projective Test was the first criterion instrument used to assess the validity of the EPPS. The RPT scores of the 30 nursing students who had been selected by the above procedures were obtained. At the time of administration of the RPT *Ss* had 19 weeks of class experience and 11 weeks of ward experience.

Presentation of material was by slides and *Ss* were told that they would be looking at situations in which nursing students might find themselves while working in the hospital. Standard TAT instructions were given and *Ss* were asked to write their stories. Each slide was projected for five minutes. The instructions also stated that the *Es* were interested in finding out, in general, what nursing students did in situations like these, that they were interested in group trends and not in individual responses per se, that there were no right or wrong answers, and that all information would be treated confidentially.

The RPT stories were scored for only six EPPS needs: three of the five needs which had been found to be high for the total group (nurturance, intraception and abasement) and three of the five needs which had been found to be low for the total group (dominance, autonomy and aggression) were scored for presence or absence. Two independent scorers achieved at least an 85 percent

agreement in scoring these protocols by means of a system based on Edwards' description of the EPPS needs. In each case, the need was scored present if it was expressed by or attributed to the nursing student depicted in the slide.

The Interaction Situation was the second criterion instrument used to assess EPPS validity. A nursing attitude questionnaire was constructed consisting of 70 items which represented issues in nursing practice and education that would arouse disagreement between nursing students.

Randomly chosen pairs of *Ss* (one from Group I and one from Group II) were tested after approximately 38 weeks of class experience, and 30 weeks of ward experience. Since there were 15 *Ss* in each Group, there were 15 experimental sessions. Subjects were individually administered the questionnaire and the first 15 items on which there was disagreement were recorded. Each *S* was then told that she would be taking part in discussions with another *S* who was in the adjacent room. Before each discussion *Ss* were to identify the statement to be discussed by reading aloud the number of the statement, the statement itself, and their responses (agree or disagree). *Ss* were told to discuss each item until they could come to a single answer for the two of them, or, if that proved impossible, until they reached some other solution. There was no time limit. The *Ss* were told that for purposes of analysis their discussions were to be recorded.

The tape recorded discussions were scored independently by two *Es* using the system outlined by Bales (1950). The scorers did not know whether a *S* belonged to Group I or Group II. Product-moment correlations between the scores of the two *Es* varied between .92 and .99 and all were significant at the .01 level.

In order to determine which Bales categories should be used to represent each of the five high and five low EPPS needs, 40 judges (graduate students in psychology) were asked to group the 12 definitions of Bales categories with the 15 definitions of EPPS needs. A "logical matching" was thus achieved. Those Bales categories which the greatest number of judges matched with each of the five high and the five low EPPS needs were treated as the need equivalents of those EPPS needs in the subsequent analysis of the IS data (See Table 1).

Separate analyses were done on the groups of three high (intra-

ception, abasement, and nurturance) and three low (autonomy, dominance, and aggression) need equivalents (which were scored across all the criterion instruments) and on the group of five high and five low need—equivalents (upon which the groups were selected from the original EPPS data).

TABLE 1
Bales Categories Matched with EPPS Needs

EPPS Need	Bales Category	Percent Agreement (<i>N</i> = 40)
Aut L*	10. Disagrees	46
Dom L	4. Gives suggestion	51
Agg L	12. Shows antagonism	86
Ach L	5. Gives opinion	33
Ord L	4. Gives suggestion	25
Int H*	8. Asks for opinion	45
Abs H	11. Shows tension	66
Nur H	1. Shows solidarity	70
Chg H	2. Shows tension release	29
Alf H	1. Shows solidarity	82

* L = Low need, H = High need.

In order to adjust the data for differences between *Ss* in terms of total number of responses made (some *Ss* gave more responses overall than others) the IS responses of each girl on each need-equivalent ("need") were transformed into a percentage of her total number of responses. In this way, the data were made ip-
native.

The Role Behavior Test was the third criterion instrument used to assess EPPS validity. For this instrument, role behavior referred to a partially structured interpersonal situation—a critical incident in nursing practice, involving a nursing student and another hospital professional or patient who interacted with her using a semi-standardized set of verbal cues. The area of interest was that of the nurse's interpersonal behavior in nursing situations scored in terms of Edward's need definitions.

The same needs which were scored on the RPT were also scored in terms of observable behavior, on the RBT.

Each critical incident contained 12 standard or set stimulus conditions to which *S* could respond. Provision was made for each of the six EPPS needs to appear twice during the four critical incidents, but never twice in the same incident. Each need segment lasted approximately six minutes.

The four roles, played by actors who were unknown to the *Ss* were: an older, irritable, worried male patient who was to have a back operation the next day, a fourth year medical student who tried to date the student nurse, an instructor in clinical nursing—particularly efficient and firm, though friendly on occasion, who conducted a self-evaluation session with the student, and a peer—an upset, somewhat hysterical young freshman nursing student who had disregarded major dormitory rules.

The four critical incident scripts contained 12 simple, direct, verbal cues which the actors memorized. The actor was permitted to ad lib (i.e. repeat or restate the original cues) if necessary to keep *S* talking and interacting, provided that he remained within the time limits specified for each stimulus condition. At the appropriate time the actor could, and did, interrupt *S* to present the next cue.

Each of the 30 *Ss* was scheduled individually after approximately 38 weeks of class experience, and 30 weeks of ward experience. A standard explanation was given indicating that *E's* concern was with groups of nursing students and not with any individual's performance in itself. The *S* was told that she would be placed in typical nursing situations which would all occur in the same room, and she was encouraged to be spontaneous. A chest microphone was placed around *S's* neck and she was told that what she said would be recorded.

The rationale for the behavioral measures used rested upon the decision to use only observable behaviors deriving directly from the six EPPS needs, and not inferences about behavior. When such behaviors were not directly applicable to the nursing incidents of this study, they were translated into nursing terminology. A validity check on these translations was made by asking ten independent raters (clinical psychologists and graduate students in clinical psychology) to match the translated behavioral statement with one of the six EPPS need categories. Only those translations on which 9 of 10 raters agreed were used as behavioral measures of an EPPS need in the four nursing role behavior situations. For example, one of the "translations" into nursing terminology for EPPS need autonomy was "to question, express doubts about staff, hospital policy, or decisions made."

A scale for duration, intensity, and frequency of responses was

used since frequency of occurrence alone gave insufficient information about the response, e.g., there is a difference between a response in which *S* touches a patient or a peer momentarily to indicate her interaction with him, and a response in which *S* touches and leans over a patient or peer for a long period of time while interacting with him. To include these differences, a 3-point scale was adopted for scoring all 17 "behaviors" representative of the six EPPS needs employed.

The tape recorded interactions between *Ss* and actors, together with the comments made by *E* which were recorded on the second track of the tape, were scored by two independent scorers in terms of the above system. Product-moment correlations between the scores of the two *Es* varied between .83 and .90, and all were significant at the .01 level.

In order to adjust the data for differences between *Ss* in terms of total number of responses made (some *Ss* gave more responses overall than others) the responses of each girl on each RBT "need" were transformed into a percentage of her total number of responses. In this way the data were made ipsative.

Hypotheses

It was predicted:

1. That the *combined* scores of Group I (Hi Hi—Lo Lo) *Ss* on the *five high* IS need-equivalents (intraception, abasement, nurturance, change, affiliation) would be significantly higher than the combined scores of Group II (Hi Lo—Lo Hi) *Ss* on those need-equivalents.

- a. That the *combined* scores of Group I (Hi Hi—Lo Lo) *Ss* on the *three high* RPT need-equivalents (intraception, abasement, nurturance) would be significantly higher than the *combined* scores of Group II (Hi Lo—Lo Hi) *Ss* on those need-equivalents.

- b. The same prediction was made for the *three RBT combined high* scores and for the *three IS combined high* scores.

2. The *combined* scores of Group I (Hi Hi—Lo Lo) *Ss* on the *five low* IS need-equivalents (autonomy, dominance, aggression, achievement, order) would be significantly lower than the combined scores of Group II (Hi Lo—Lo Hi) *Ss* on those need-equivalents.

- a. That the *combined* scores of Group I (Hi Hi—Lo Lo) *Ss*

on the *three low* RPT, need-equivalents (autonomy, dominance, aggression) would be significantly lower than the combined scores of Group II (Hi Lo—Lo Hi) Ss on those need-equivalents.

b. The same prediction was made for the *three* RBT combined low scores and for the *three* IS combined low scores.

3. That the scores of the Group I (Hi Hi—Lo Lo) Ss on each of the *five high* IS need-equivalents would be significantly higher than the scores of the Group II (Hi Lo—Lo Hi) Ss on each of those need-equivalents.

a. That the scores of the Group I (Hi Hi—Lo Lo) Ss on each of the *three high* RPT need-equivalents would be significantly higher than the scores of the Group II (Hi Lo—Lo Hi) Ss on each of those need-equivalents.

b. The same prediction was made for each of the *three high* RBT need-equivalents.

4. That the scores of the Group I (Hi Hi—Lo Lo) Ss on each of the *five low* IS need-equivalents would be significantly lower than the scores of the Group II (Hi Lo—Lo Hi) Ss on each of those need-equivalents.

a. That the scores of the Group I (Hi Hi—Lo Lo) Ss on each of the *three low* RPT need-equivalents would be significantly lower than the scores of the Group II (Hi Lo—Lo Hi) Ss on each of those need-equivalents.

b. The same prediction was made for each of the *three low* RBT need-equivalents.

5. That the scores of the total group of 30 Ss on each EPPS need studied would correlate higher with the same Ss' scores on each corresponding RPT need-equivalent than with any other RPT need-equivalent.

a. The same prediction was made for each EPPS need studied and each of the RBT and IS need-equivalents.

Results

Testing Validity in Terms of Group Differences

It was predicted that Groups I and II, differentiated on the five high and five low needs of the EPPS, should also be differentiated on the five high and five low (or three high and three low) EPPS need-equivalents of the RPT, the IS, and the RBT. In evaluating

TABLE 2

Differences between Group I and Group II on Role Projective Test (RPT), Interaction Study (IS), and Role Behavior Test (RBT) Data

RPT "need"	G I-Frequency of "presence" scores ^b	G II-Frequency of "presence" scores ^b	χ^2
Aut L ^a	1	4	0.99
Dom L	1	5	1.95
Agg L	0	0	0
Int H	9	9	0
Aba H	4	4	0
Nur H	5	5	0
	G I- \bar{f} of Ss above median	G II- \bar{f} of Ss above median	χ^2
Combined "presence" score: 3 low needs	6	2	1.63
Combined "presence" score: 3 high needs	5	7	0.62
IS "Need"	G I Mean Score	G II Mean Score	t, χ^2 , or U
Aut L	6.29	7.41	$t = 1.75^*$
Dom L	1.30	2.00	$t = 1.84^*$
Agg L	2 ^b	3 ^b	$\chi^2 = 0$
Ach L	30.80	26.80	—
Ord L	1.30	2.00	$t = 1.84^*$
Int H	1.80	2.30	—
Aba H	15.01	14.39	$t = 0.63$
Nur H	1.47	1.63	—
Chg H	2.80	3.40	—
Aff H	1.47	1.63	—
Sum ranks for 5 low needs ^a	1177.0	1010.5	U = 82
Sum ranks for 5 high needs	1120.5	1137.0	U = 92
Sum ranks for 3 low needs	597.0	455.0	U = 47 [*]
Sum ranks for 3 high needs	534.0	520.0	—
RBT "Need"	G I Mean Score	G II Mean Score	t or U
Aut L	6.54	10.98	$t = 2.01^*$
Dom L	11.31	14.67	$t = 0.94$
Agg L	5.08	5.86	$t = 0.34$
Int H	19.34	17.52	$t = 0.65$
Aba H	14.62	13.00	$t = 0.44$
Nur H	43.11	38.15	$t = 1.26$
Sum ranks for 3 low needs	612.5	440.5	U = 49 [*]
Sum ranks for 3 high needs	478.5	583.5	U = 55

* Significant at .05 level (one tailed).

^a Prediction throughout was that G I (Hi Hi-Lo Lo) would score higher (or show greater frequency) on high needs (H) and lower (or show greater frequency) on low (L) needs than G II (Hi Lo-Lo Hi).

^b Number of S's (of 15) giving such response.

^c Smaller sum rank signifies larger need score throughout.

these hypotheses, tests of the differences were made between Groups I and II on each EPPS need-equivalent, as well as on the combined "high" need-equivalents and the combined "low" need-equivalents.

In general, EPPS needs did not prove to be valid in terms of the criteria employed. Only one need, autonomy, was found to differentiate the groups on more than one criterion (IS and RBT).

Role Projective Test

Comparisons between Group I and Group II involved three low need-equivalents (autonomy, dominance, aggression) and three high need-equivalents (intraception, abasement, nurturance). The group differences on these need-equivalents were assessed by chi square, since the data were nominal, i.e. a *S*'s protocol was scored for the "presence" of a need if it appeared in any of her stories which were scored for that need ("presence" scores in Table 2). Results are presented in Table 2. There were no significant differences between the groups on any RPT need-equivalent. In addition, only two need-equivalents (autonomy, dominance) yielded trends in the predicted direction.

In order to obtain combined "presence" scores for the groups on the three low need-equivalents, *Ss*' "presence" scores were first summed across the three low need-equivalents. The median of this distribution was then obtained as well as the number of *Ss* above and below the median in each group. Chi square was then computed to assess the differences between the two groups. Although there was a trend in the predicted direction, it was not significant. The same method was used to obtain the combined "presence" scores for the groups on the three high need-equivalents. In this case, the trend was not in the predicted direction.

Interaction Study

Two sets of comparisons between Group I and Group II on this criterion instrument were made, the first involving all five low need-equivalents (autonomy, dominance, aggression, achievement, order) and all five high need-equivalents (intraception, abasement, nurturance, change, affiliation) and the second involving only the three high and three low need-equivalents scored for the other two criterion measures. Differences between the groups on

these need-equivalents were assessed by t -tests, except in the case of need-equivalent aggression. In this case the group difference was assessed by chi square, since only three Ss in Group I and two Ss in Group II had scores on IS need-equivalent aggression. As can be seen from Table 2, the groups were significantly different in the predicted direction on three of the individual IS need-equivalents, although this meant that they were significantly different on only two of the Bales categories (since Bales category 4 had been judged as logically matching both EPPS order and dominance—see Table 1). In addition, one trend was in the predicted direction but not statistically significant—that for IS need-equivalent abasement.

In order to compare the two groups on the sum of the low need-equivalents, all Ss' scores on each of the five low need-equivalents were ranked, and then these ranks were summed across the five low need-equivalents. A U-test of the significance of the differences between Group I and II on these sum ranks was then computed. The same procedure was used to compare the two groups on the sum of the five high need-equivalents, and on the sum of the three low and the sum of three high need-equivalents. Of the comparisons between the groups on these sum ranks, only one, that for the combined three low IS need-equivalents, was both significant and in the predicted direction. Group differences on the sum ranks of the five low and five high need-equivalents were in the predicted direction but not significant, and the difference between the groups on the sum ranks of the three high need-equivalents was not in the predicted direction.

Role Behavior Test

Comparisons between Group I and Group II on this criterion instrument involved the same three low and three high need-equivalents which were assessed both on the RPT and the IS. For the RBT, the data permitted the use of t tests to assess the differences between groups. The results of these tests are presented in Table 2. Of the comparisons between the groups on individual need-equivalents, only RBT need-equivalent autonomy significantly differentiated the groups in the predicted direction. Although not significant, the trends for all other RBT need-equivalents were in the predicted direction.

The same method as that used to obtain the sums of the low and

high need-equivalents for the IS was also used with the RBT scores. For the RBT, U-tests revealed that the sum ranks of the three low need-equivalents significantly differentiated the groups in the predicted direction, while the group difference on the sum ranks of the three high need-equivalents was in the predicted direction, but not statistically significant.

In general, the equivalents of EPPS need autonomy were the only measures out of the six need-equivalents assessed on all criterion instruments which significantly differentiated the groups on more than one of those instruments (IS and RBT). Since this was a low need in the total nursing group it is possible that a need which is low in the group being assessed is able to provide the greatest discrimination. The only other need-equivalents to significantly differentiate the groups were IS need-equivalents dominance and order. Since these were both related to the same Bales category, however, they refer to a common set of scores.

Testing Validity in Terms of Correlation

If a measure of a need or trait on a particular instrument correlates higher with corresponding measures of that need or trait than with measures of any other need or trait on other instruments, then support is obtained for the assumption that the original instrument is a valid measuring device for the need or trait in question. Accordingly, in the present investigation correlations were obtained between certain EPPS needs and equivalent needs measured on the criterion instruments (RPT, IS, RBT), the prediction being that there would be a higher positive correlation between an EPPS need and its equivalent (same-name) need on the other instruments than between an EPPS need and any non-equivalent (other-name) need measured on the instruments.

Role Projective Test

Point-biserial intercorrelations were obtained between the EPPS needs being investigated and their RPT need-equivalents. It was predicted, for example, that the scores for the 30 Ss on EPPS need autonomy should correlate most positively with the RPT scores for autonomy but not with the RPT scores for any of the other needs. It was not possible to include the need aggression in the intercorrelations because there were no RPT scores on this need-

equivalent. Table 3 presents the correlation between each EPPS need and its RPT need-equivalent, and also the RPT need-equivalent which actually showed the highest positive correlation with each EPPS need being considered. Thus, in Table 3, the RPT measure for need autonomy correlated $+ .12$ with the actual EPPS scores of *Ss* on this need. The RPT measure for need autonomy was also found to be the RPT measure which showed the highest correlation with the actual EPPS scores for need autonomy. The RPT

TABLE 3
Correlations between EPPS Needs and RPT, IS, and RBT Data

RPT measure corresponding to EPPS need	r_{pb}	RPT measure correlating highest with EPPS need	r_{pb}
Aut L ^b	$+ .12$	Aut	$+ .12^*$
Dom L	$+ .39$	Aut	$+ .61$
Agg L	—	No scores	—
Int H ^b	$- .05$	Nur	$+ .20$
Aba H	$+ .02$	Nur	$+ .16$
Nur H	$+ .07$	Aba	$+ .17$
IS measure corresponding to EPPS need	r	IS measure correlating highest with EPPS need	r
Aut L	$+ .17$	Nur/Aff	$+ .43$
Dom L	$+ .27$	Dom/Ord	$+ .27^*$
Agg L	$+ .60$	Agg	$+ .60^*$
Ach L	$- .11$	Chg	$+ .37$
Ord L	$+ .27$	Ord/Dom	$+ .27^*$
Int H	$- .16$	Ach	$+ .19$
Aba H	$+ .05$	Int	$+ .52$
Nur H	$- .35$	Ach	$+ .43$
Chg H	$- .23$	Aba	$+ .38$
Aff H	$- .06$	Ach	$+ .26$
RBT measure corresponding to EPPS need	r	RBT measure correlating highest with EPPS need	r
Aut L	$+ .31$	Dom	$+ .34$
Dom L	$+ .16$	Aut	$+ .27$
Agg L	$+ .53$	Agg	$+ .53^*$
Int H	$- .02$	Aba	$+ .08$
Aba H	$- .01$	Nur	$+ .32$
Nur H	$+ .09$	Aba	$+ .12$

* See Table 1 for corresponding Bales categories.

^b L = Low need; H = High need.

* Results conform to prediction.

measure for need dominance correlated $+.39$ with the actual EPPS scores for need dominance. However, it was found that the RPT measure for need autonomy correlated higher with the actual EPPS scores for need dominance ($+.61$) than did the RPT measure for need dominance.

It can be seen that the prediction was borne out only for need autonomy. Further analysis of the correlations was not attempted, since, although it appears reasonable to compare the relative magnitudes of correlation for this sample, it is probably not reasonable to offer any conclusions about their statistical significance. Tests of significance imply random samples of *Ss*, and the present groups were not selected in this manner.

Interaction Situation

The scores of the 30 *Ss* on the EPPS needs selected for study were correlated by means of Pearson r with the scores of these *Ss* on the Bales categories to which the EPPS needs had been logically matched. As can be seen from Table 3, the prediction that logically corresponding needs should correlate most positively was borne out for three needs: order, dominance, and aggression. In the cases of EPPS needs order and dominance, however, the result is somewhat spurious, since both had been matched to the same Bales category (#4). Doubt must be cast on the reliability of the aggression correlation since in the IS only five *Ss* obtained scores on the logically matched Bales category (#12). It is clear then that using the IS criterion there is little support for the hypothesis that EPPS need scores are valid measures of behavior.

Role Behavior Test

Pearson product-moment correlations were computed between *Ss'* EPPS scores and their scores on the six needs which were assessed on the RBT. The correlations between same-name (equivalent) needs and also the correlations between each EPPS need and the highest correlating RBT need-equivalent are presented in Table 3. Only one correlation, for need aggression, conformed to the prediction that same-name needs should correlate most positively. In two cases, for EPPS needs abasement and intraception, the correlations with RBT need-equivalents were negative. These findings suggest that except for EPPS aggression there was little relation-

ship between what was being measured on the EPPS and on the RBT.

Intercorrelations between the EPPS needs and the RPT, IS, and RBT need-equivalents were also performed by means of the correlation ratio (η). Again, no consistent findings in the directions predicted by the hypotheses were obtained. Correlational approaches to the assessment of EPPS validity did not therefore lend any consistent support to the validity of EPPS needs.

Discussion

Neither of the two approaches to the testing of validity adopted in the present study (testing of group differences and correlational analysis) resulted in findings which consistently supported the adequacy of the EPPS as a measure of certain personality needs. These findings are consistent with the findings of Fisher and Morton (1957) and of Endler (1961) that the EPPS was not sensitive to known differences between groups. Some consideration, however, should be given to the possibility that EPPS need autonomy is a valid measure. In this connection, it is of interest that need autonomy was also found by Zuckerman (1958) to be capable of differentiating a "rebellious" group from three "dependency" groups.

Sheldon, Coale, and Copple (1959) found that certain EPPS needs were sensitive to group differences based on MMPI scores. Four of the five needs which Sheldon et al. found differentiated the MMPI-selected groups (aggression, dominance, abasement, and affiliation) were studied in the present investigation.

A possible explanation of the discrepancy in results between the Sheldon et al. (1959) study and the present investigation lies in the fact that the former investigators made predictions from six specific EPPS needs which were chosen because they were rationally considered most likely to differentiate between high and low scoring groups on the Minnesota Teacher Attitude Inventory and the MMPI. In the present investigation, on the other hand, choice of needs was determined on a purely empirical basis. In addition, although it was required that all Ss selected for Groups I and II had to obtain scores on the five high and five low group needs which were more extreme than the means of the original group, there was no requirement that they had to deviate on all, or the same, needs. As long as their total deviation score was sufficiently

large, *Ss* were included in either Group I or Group II. Although Groups I and II were significantly different on all the needs studied some overlap of scores on each of the needs did occur. This may account in part for the failure to obtain significant separation between the groups on the criterion measures.

In the interaction situation (IS), it is evident (as is shown in Table 1) that interjudge agreement was not high for some of the matchings of Bales categories and EPPS needs. Low agreement would of course influence IS measures of validity. Another possible failing of the present study is that the other measures employed may not have been assessing the same characteristics as those assessed by the EPPS. Although this latter possibility cannot be excluded it seems unlikely that (a) an unstructured projective test, (b) a series of behavioral situations specifically designed to elicit behaviors similar to those which Edwards proposed his needs were measuring, and (c) a discussion situation which gave *Ss* considerable scope to express their personality characteristics, would fail to register aspects of personality in groups which were differentiated on the EPPS.

If the factor of social desirability constitutes an important source of variance on the EPPS, as stated by Corah et al. (1958), it too could account in part for the discrepancy between the scores of *Ss* on the EPPS as opposed to their scores on the other measures. Social desirability is not likely to have had as much influence on *Ss'* responses to the projective and behavioral measures employed as it might have had on their EPPS responses. On the EPPS, in addition, *Ss* were presumably responding in terms of self-appraisal, whereas this variable was probably of much less importance in influencing their projective or behavioral responses.

Examining the criterion measures used, it appears that the behavioral instruments were more sensitive to group differences than was the projective instrument. This conforms to the idea that real behavior is the best indicator of personality, but, on the other hand, it does not support the alternative idea that a criterion instrument which is "closer" to the original instrument should show more agreement with it than criterion instruments which differ more from the original instrument. Presumably in the present investigation the projective measure (RPT) was more similar to the EPPS than were the behavioral measures employed in that in-

teraction with another person was not elicited by the EPPS or the RPT.

Summary

The validity of the EPPS was assessed using nursing student Ss. Two groups were differentiated on the basis of EPPS scores and then tested on three behavioral and projective measures. It was predicted that differences between groups on the EPPS should also be found on the other measures, and the EPPS need scores should show higher correlations with corresponding than with non-corresponding scores on the other measures. Essentially negative findings cast doubt on the validity of the EPPS, although some support was obtained for need autonomy. Methodological problems were discussed.

REFERENCES

- Allen, R. M. "The Relationship between the Edwards Personal Preference Schedule Variables and the Minnesota Multiphasic Personality Inventory Scales." *Journal of Applied Psychology*, XLI (1957), 307-311.
- Bales, R. F. *Interaction Process Analysis*. Cambridge, Mass.: Addison-Wesley, 1950.
- Barber, W. H. and Wessen, A. F. "Perspective and Strategy for Nursing Role Research." *Nursing Research*, in press.
- Bendig, A. W. "Objective Measures of Needs and Course Achievement in Introductory Psychology." *Journal of General Psychology*, LIX (1958), 51-57.
- Bernardin, A. C. and Jessor, R. "A Construct Validation of the EPPS with Respect to Dependency." *Journal of Consulting Psychology*, XXI (1957), 63-67.
- Corah, N. L., Feldman, M. J., Cohen, I. S., Gruen, W., Meadow, A., and Ringwall, E. A. "Social Desirability as a Variable in the Edwards Personal Preference Schedule." *Journal of Consulting Psychology*, XXII (1958), 70-72.
- Dunette, M. D., Kirchner, W. K., and De Gidio, Jo Anne. "Relations among Scores on Edwards Personal Preference Schedule, California Psychological Inventory, and Strong Vocational Interest Blank for an Industrial Sample." *Journal of Applied Psychology*, XLII (1958), 178-181.
- Edwards, A. L. *Manual for the Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1959.
- Endler, N. S. "Conformity Analyzed and Related to Personality." *Journal of Social Psychology*, LIII (1961), 271-283.
- Fisher, S. and Morton, R. B. "An Exploratory Study of Some Relationships between Hospital Ward Atmospheres and Atti-

- tudes of Ward Personnel." *Journal of Psychology*, XLIV (1957), 155-164.
- Gebhart, G. G. and Hoyt, D. P. "Personality Needs of Under- and Over-achieving Freshmen." *Journal of Applied Psychology*, XLII (1958), 125-128.
- Goodstein, L. D. and Heilbrun, A. B., Jr. "Prediction of College Achievement from the Edwards Personal Preference Schedule at Three Levels of Intellectual Ability." *Journal of Applied Psychology*, XLVI (1962), 317-320.
- Graine, G. N. "Measures of Conformity as Found in the Rosenzweig P-F Study and the Edwards Personal Preference Schedule." *Journal of Consulting Psychology*, XXI (1957), 300.
- Hartley, R. E. and Allen, R. M. "The Minnesota Multiphasic Personality Inventory (MMPI) and the Edwards Personal Preference Schedule (EPPS): A Factor Analytic Study." *Journal of Social Psychology*, LVIII (1962), 153-162.
- Heilbrun, A. B., Jr. "Relationships between the Adjective Checklist, Personal Preference Schedule, and Desirability Factors under Varying Defensiveness Conditions." *Journal of Clinical Psychology*, XIV (1958), 283-287.
- Heilbrun, A. B., Jr. and Goodstein, L. D. "Relationships between Personal and Social Desirability Sets and Performance on the Edwards Personal Preference Schedule." *Journal of Applied Psychology*, XLIII (1959), 302-305.
- Heilbrun, A. B., Jr. and Goodstein, L. D. "Social Desirability Response Set: Error or Predictor Variable." *Journal of Psychology*, LI (1961), 321-329.
- Izard, C. E. "Personality Characteristics of Engineers as Measured by the Edwards Personal Preference Schedule." *Journal of Applied Psychology*, XLIV (1960), 332-335.
- Klett, C. J. "The Stability of the Social Desirability Scale Values in the Edwards Personal Preference Schedule." *Journal of Consulting Psychology*, XXI (1957), 183-185. (a)
- Klett, C. J. "The Social Desirability Stereotype in a Hospital Population." *Journal of Consulting Psychology*, XXI (1957), 419-421. (b)
- Klett, C. J. and Tamkin, A. S. "The Social Desirability Stereotype and Some Measures of Psychopathology." *Journal of Consulting Psychology*, XXI (1957), 450.
- Klett, C. J. and Yaukey, D. W. "A Cross-cultural Comparison of Judgments of Social Desirability." *Journal of Social Psychology*, XLIX (1959), 19-26.
- Krug, R. E. "Over- and Under-achievement and the Edwards Personal Preference Schedule." *Journal of Applied Psychology*, XLIII (1959), 133-136.
- Krug, R. E. and Moyer, K. E. "An Analysis of the F Scale: II. Relationship to Standardized Personality Inventories." *Journal of Social Psychology*, LIII (1961), 293-301.
- Lang, G., Sferra, A. G., and Seymour, M. "Psychological Needs of

- College Freshmen and Their Academic Achievement." *Personnel Guidance Journal*, XLI (1962), 359-360.
- Marlowe, D. "Relationships among Direct and Indirect Measures of the Achievement Motive and Overt Behavior." *Journal of Consulting Psychology*, XXIII (1959), 329-332.
- Mogar, R. E. "Competition, Achievement, and Personality." *Journal of Counseling Psychology*, IX (1962), 168-172.
- Sheldon, M. S., Coale, J. M., and Copple, R. "Concurrent Validity of the 'Warm Teacher Scales.'" *Journal of Educational Psychology*, L (1959), 37-40.
- Worell, L. "EPPS N Achievement and Verbal Paired-Associates Learning." *Journal of Abnormal and Social Psychology*, LX (1960), 147-150.
- Zuckerman, M. "The Validity of the Edwards Personal Preference Schedule in the Measurement of Dependency-Rebelliousness." *Journal of Clinical Psychology*, XIV (1958), 379-382.
- Zuckerman, M. and Grosz, H. J. "Suggestibility and Dependency." *Journal of Consulting Psychology*, XXII (1958), 328.
- Zuckerman, M., Levitt, E. E., and Lubin, B. "Concurrent and Construct Validity of Direct and Indirect Measures of Dependency." *Journal of Consulting Psychology*, XXV (1961), 316-323.

VOCATIONAL PREFERENCE PATTERNS OF COMMUNICATIONS GRADUATES*

ALLEN E. IVEY

Colorado State University

AND

MARK B. PETERSON

Boston University

THE communications fields of public relations, journalism, and radio-television attract considerable attention and interest among young people nearing a vocational choice. There is relatively little information available on the vocational interest patterns of those who complete the educational requirements for employment in these vocations. This study is designed to supply some preliminary data on the vocational interest patterns of communications graduates. In addition, this study examines the capability of an interest test, the Kuder Preference Record—Vocational (KPR-V), to differentiate among apparently closely related sub-groups in communications fields.

Cranford (1960) and Weigle (1957) using interview techniques ascertained that decisions to enter journalism were made during high school. Fosdick (1961) asked high school students involved in extracurricular journalism activities to rate nine occupations on five different dimensions and found a high level of "interest" in journalism. The findings of Fosdick's study correspond closely with an earlier study of attitudes by Lubell (1959). Data are also available from the Kuder Administrator's Manual (1960) on

* The authors wish to acknowledge the aid of Professor John Hale, Bucknell University Computational Laboratory, for handling the machine computation of the data in this study.

"Writers" and "Advertising Personnel." Although the characteristics of the samples are not specified, high literary patterns are noted for both groups. A review of the literature revealed no direct information on the interest patterns of radio-television or public relations personnel.

Research has shown that the KPR-V can successfully differentiate among occupations (Kuder, 1956; Pierce-Jones, 1959). A more difficult problem occurs when the counselor has to help the student make a choice among closely related jobs in the same occupational area. Studies such as those by Baas (1950), Brody (1957), Reed, Lewis, and Wolins (1960), and Triggs (1948) on psychologists, foresters, engineers, and nurses have shown that the KPR-V can discriminate among sub-types of occupations. Sternberg (1955) has shown that curriculum choice can be differentiated by the KPR-V.

While we have evidence that the KPR-V can differentiate among occupations and among sub-types within occupational areas, relatively little information is available on the interest patterns of communications graduates. Counseling situations frequently arise where more complete information on the fields of radio-television, journalism, and public relations would be helpful. This study is designed to provide preliminary data on the interest patterns of graduates in communications fields and also to consider the ability of the KPR-V to differentiate among the three related fields of radio-television, journalism, and public relations.

Method

The subjects of this investigation were 108 male graduates of the Boston University School of Public Relations and Communications during the four year period 1958 to 1961. Forty of these students had graduated in Communications Arts (radio-television), 23 in Journalism, and 45 in Public Relations.

All of the students involved in this study had taken their first two years at Boston University in the College of General Education, a two-year integrated liberal studies program, prior to transferring to the School of Public Relations and Communications. In their first two years all students take the same required courses which makes their school experience quite similar (Walston, 1960). In the last two years at the School of Public Relations and

Communications there is specialization in the three subject matter fields of radio-television, public relations, and journalism.

In September of the freshman year (1954-1957), the students were given the KPR-V as a portion of the Guidance Department's freshman orientation program. It is these scores which were examined to determine the KPR-V patterns of communications graduates four years later. The questions being considered were: 1) "What are the KPR-V patterns of communications graduates?"; and 2) "Can data taken during the freshman year be used to discriminate effectively among graduates four years later in the three communications areas?"

Preliminary analysis of the data involved computation of the means for each of the ten KPR-V scales. An analysis of variance was then employed on each scale to determine if there were significant variations among the three groups. The nature of any significant differences was examined further by the *t* test. It is recognized that the KPR-V scales are partially ipsative and that multivariate analysis might be considered a more refined statistical technique. However, due to the preliminary nature of this study, analysis of variance and *t* tests were employed. From this analysis, future directions for investigation may be suggested.

Results and Discussion

The mean raw scores along with their corresponding percentiles, and resultant *F*'s for each of the three groups on the ten KPR-V scales are listed in Table 1.

An analysis of the mean scores converted to percentile scores reveals that all three groups were above the 74th percentile on the persuasive scale, above the 80th percentile on the literary scale, and above the 65th percentile on the musical scale. It may be further seen that the three groups tend to score in the lower ranges of the outdoor, mechanical, computational, and scientific scales. In general, one might be impressed by the similarity of the KPR-V patterns of students who later have gone into radio-television, journalism, or public relations. These results seem to support the information presented in the latest edition of the Administrator's Manual (Kuder, 1960).

However, it is also interesting to note that communications graduates did demonstrate differences on the four scales listed in Table

TABLE 1
Mean Raw Scores, Corresponding Percentiles, and F's for Communications Graduates on the Kuder Preference Record—Vocational

Curriculum						
Scale	Radio-Television (<i>N</i> = 40)		Journalism (<i>N</i> = 23)		Public Relations (<i>N</i> = 45)	
	Mean Raw Score	%ile Rank of Mean Raw Score	Mean Raw Score	%ile Rank of Mean Raw Score	Mean Raw Score	%ile Rank of Mean Raw Score
Outdoor	30.5	21	29.6	20	28.8	19
Mechanical	28.2	13	25.3	10	31.3	17
Computational	19.4	16	23.6	30	21.6	23
Scientific	31.8	24	28.0	14	33.1	28
Persuasive	50.7	74	53.7	79	59.3	87
Artistic	25.3	63	20.1	41	20.9	45
Literary	28.9	84	36.5	98	27.4	80
Musical	21.3	88	14.6	65	16.1	73
Social Service	34.8	30	38.6	40	38.4	40
Clerical	46.5	55	53.0	72	46.9	57
						<i>F</i>
						0.19
						1.65
						2.08
						1.70
						5.27**
						3.94*
						12.38**
						9.29**
						1.01
						2.44

* 0.05 level of significance.

**** 0.01 level of significance.**

1. Significant F ratios at the 0.01 level of significance exist on the persuasive, literary, and musical scales. The significance of the difference on the artistic scale was at the 0.05 level. Despite similarities in profile patterns, it does seem that the KPR-V reveals differences among communications graduates.

Table 2 presents inter-curricula comparisons of four KPR-V scales and allows for further analysis of the data.

TABLE 2
Inter-curricula Comparisons of Four KPR-V Scales

Curriculum Comparison	t -value			
	Persuasive	Artistic	Literary	Musical
Radio-Television vs. Journalism	0.966	2.694**	5.100**	3.585**
Radio-Television vs. Public Relations	3.039**	2.245*	0.872	2.012*
Journalism vs. Public Relations	2.057*	0.514	6.105**	0.836

* 0.05 level of significance.

** 0.01 level of significance.

The results of the t tests reveal what would logically be anticipated for the different curricula. Graduates in the field of radio-television had significantly higher scores on the musical and artistic scales than did public relations or journalism majors. Journalism majors had significantly higher literary interests than public relations or communications students. Finally, the public relations graduates had significantly higher interests on the persuasive scale than did those in journalism or radio-television. Since the t -values were computed following the computation of the analysis of variance, they should not be considered as conclusive, but as supportive of hypotheses to be tested in subsequent studies.

As Reed, Lewis, and Wolins (1960) have noted, knowledge of interest patterns may have influenced the curriculum choice of students. The students in this study had been given information concerning their KPR-V patterns. One may only hypothesize about the degree of relationship between this knowledge and later curriculum choice. It is suggested that the general similarity of interest patterns among radio-television, journalism, and public relations

students may partially "mask" the influence of test interpretations. In addition, the counselors who interpreted the results of the KPR-V to students had no knowledge that this interest inventory could differentiate among the groups.

It must be recognized that the students involved in this study came from one college in one university. The sample is not large, although it did include all of those four graduating classes in communications who had attended the College of General Education. The findings of this study, as such, must be interpreted with some caution.

Lending significance to the data is the fact that the communications students were tested in their freshman year, which was two years prior to their committing themselves to a specific communications area. Further, many of the students involved in this study had no definite curriculum choice at the time they took the KPR-V. It is noteworthy that the KPR-V can differentiate among these closely related curricula at an early time in a student's college career. Also contributing to the meaningfulness of the data is the fact that the test scores were obtained over a four-year period and interpretations involved several different counselors.

Counseling implications of these data should be considered. First, it would seem that the KPR-V does identify certain general characteristics of individuals who may wish to consider careers in the communications field. Once a commitment or interest has been expressed by a student in the communications field, it does seem that the counselor could advise the student from his KPR-V profile as to which specific area or major field might be considered. However, while the KPR-V does appear to distinguish between these curricula at an early stage, it must be recognized that the practical spread of the differences for counseling purposes is not large, and that there is overlap between the three fields. Further, due to statistical limitations, these results should be considered suggestive.

Summary

This study was designed to provide preliminary data on the interest patterns of communications graduates through an analysis of scores on the Kuder Preference Record—Vocational. The study also examined the ability of the KPR-V to differentiate among the in-

terest patterns of three sub-groups within communications: radio-television, journalism, and public relations. The following conclusions seem apparent from data collected when the students were freshmen.

1. Communications graduates tend to have high KPR-V persuasive, literary, and musical scales. They tend to have low scores on outdoor, mechanical, computational, and scientific scales.
2. It was found that radio-television, journalism, and public relations students differed among themselves on four KPR-V scales: persuasive, literary, musical, and artistic. The nature of the specific differences among these groups was also examined. Preliminary data suggests that radio-television graduates tend to have significantly higher scores on the artistic and musical scales, journalism students rank higher on the literary scale, and public relations graduates show most interest on the persuasive scale.

The findings of this study should prove helpful to the counselor aiding students considering the communications fields. However, due to limitations of the study described herein, the findings should be interpreted with some caution.

REFERENCES

- Baas, M. L. "Kuder Interest Patterns of Psychologists." *Journal of Applied Psychology*, XXXIV (1950), 115-117.
- Brody, D. S. "Kuder Interest Patterns of Professional Forest Service Men." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENTS*, XVII (1957), 599-605.
- Cranford, R. J. "When Are Career Choices for Journalism Made?" *Journalism Quarterly*, XXXVII (1960), 422-424.
- Fosdick, J. A. and Greenberg, B. S. "Journalism as a Career Choice: A Small Sample Study." *Journalism Quarterly*, XXXVIII (1961), 380-382.
- Kuder, G. F. *Examiner Manual*, Kuder Preference Record, Vocational, Form C. Chicago: Science Research Associates, 1956.
- Kuder, G. F. *Administrator's Manual*, Kuder Preference Record, Vocational Form C. Chicago: Science Research Associates, 1960.
- Lubell, S. "High School Students' Attitudes toward Journalism as a Career." *Journalism Quarterly*, XXXVI (1959), 199-203.
- Pierce-Jones, J. "Kuder Preference Record—Vocational." In O. K. Buros (Ed.), *The Fifth Mental Measurements Yearbook*. Highland Park, New Jersey: Gryphon, 1959, 891-892.
- Reed, W. R., Lewis, E. C., and Wolins, L. "Differential Interest Patterns of Engineering Graduates." *Personnel and Guidance Journal*, XXXVIII (1960), 571-573.

- Sternberg, C. "Personality Traits of College Students Majoring in Different Fields." *Psychological Monographs*, 69(18) (1955) No. 403.
- Triggs, F. O. "The Measured Interests of Nurses: A Second Report." *Journal of Educational Research*, XLII (1948), 113-121.
- Walston, E. B. "Personal and Occupational Psychology at Boston University College of General Education." In H. T. Morse and P. T. Dressel (Eds.), *General Education for Personal Maturity*. Dubuque, Iowa: W. C. Brown, 1960, 59-72.
- Weigle, C. F. "Influence of High School Journalism on Choice of Career." *Journalism Quarterly*, XXXIV (1957), 39-45.

ELECTRONIC COMPUTER PROGRAMS AND ACCOUNTING MACHINE PROCEDURES

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

<i>An IPL-V Technique for Simulation Programs.</i> FRANK B. BAKER AND THOMAS J. MARTIN	859
<i>A FORTRAN Generator of Polynomials Orthonormal over Unequally Spaced and Weighted Abscissas.</i> PHILLIP L. EMERSON	867
<i>A Subroutine to Refine the Inverse of a Matrix.</i> NATHAN JASPEN	873
<i>The Calculation of Probabilities Corresponding to Values of z, t, F, and Chi-Square.</i> NATHAN JASPEN	877
<i>Scorit—A FORTRAN Program for Scoring and Item Analysis of Porta-punch Test Cards.</i> PAUL A. GAMES	881
<i>Data Processing Procedures to Improve Classroom Testing.</i> QUENTIN C. STODOLA	885
<i>Variables Affecting the Graduate Assistant in a Computer Training Position.</i> M. GORDON HOWAT	887

In view of the tremendous advances that have been made in the adaptation of electronic computers and accounting machines to the processing of statistical data, sections of the Spring and Autumn issues of **EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT** are devoted to the publication of such programs as are appropriate to psychometric procedures. Programs relevant to such problem areas as factor analysis, item analysis, multiple regression procedures, the estimation of the reliability and validity of tests, pattern and profile analysis, the analysis of variance and covariance, discriminant analysis, and test scoring will be considered. Customarily a program should be expected not to exceed six or eight printed pages. Manuscripts of four or fewer printed pages are preferred. Each manuscript will be carefully reviewed as to its suitability and accuracy of content. In some instances an accepted paper may be returned to the author for possible revisions or shortening. The cost to the author will be fifteen dollars per page for regular running text. The extra cost of the composition of tables and formulas will be added to the basic rate. Manuscripts received up to November first will be considered for the Spring issue; manuscripts received between then and May first will be considered for the Autumn issue.

All correspondence should be directed to

William B. Michael
Professor of Education and Psychology
University of California, Santa Barbara
Santa Barbara, California

AN IPL-V TECHNIQUE FOR SIMULATION PROGRAMS

FRANK B. BAKER

AND

THOMAS J. MARTIN

University of Wisconsin

DURING the early developmental stages most simulation programs have an exceedingly short half-life. The investigator is usually conducting research of an exploratory nature and the discovery of an important relationship, realization of a significant omission, and other such events occur with alarming frequency. The insights thus obtained can yield changes in the computer program which range from complete obsolescence to a minor restructuring of the program. Having been engaged in the development of programs which simulate concept learning (Baker, 1964), the authors are painfully aware of the highly fluid nature of such programs. It was within this context that awareness of a technological problem intimately associated with simulation programming arose—namely, one must possess the capability to minimize the impact of significant changes while one is simultaneously maximizing the ability to make such changes. The present paper describes a pseudo-code system and an interpreter both using IPL-V (Newell, 1964) which the writers offer as a reasonable solution to this technological problem. Although the technique was developed within the context of a particular simulation problem, the writers feel that it is applicable to a much wider class of problems than that posed by the immediate problem.

The pseudo-code consists of an IPL-V list containing symbols representing routines which are to be performed as well as local symbols which indicate branches in the program. These lists do not,

however, contain any IPL-V primitives and the lists are not executable IPL-V programs. Table 1 consists of a list which represents a typical concept learning strategy expressed as a list of symbols.

TABLE 1
List Representing Conservative Strategy

Symbol		Description
S2	9-0	
	Z0	Obtain problem definition
9-1	Z1	Vary an attribute value
	Z2	Select an object
	Z3	Experimenter designates object
	Z4	React to designation
	D1	Determine if concept to be offered
	9-1	No
	Z5	Yes, experimenter designates concept
	D2	Determine if problem solved
	9-2	Yes
	X11	No, print problem history
9-2	0	0

Each symbol on the list can be the name of a list of symbols, and this representational form can be carried to any depth desired. These symbols are referred to as pseudo-codes, as they are merely abstract representations of a psychological process. In one learning simulation program there are three levels in the list structure which constitute the simulation program. The highest level, the *S* level, is essentially an executive-level description of the over-all learning strategy. The second level consists of major procedures—the *Z* or *D* routines—which perform salient tasks such as hypothesis generation. The third and lowest level are the *P*'s and *Q*'s which are actually executed to perform the information processing necessary for concept learning. The *P*'s and *Q*'s are contained within the *Z*'s and *D*'s, and the *Z*'s and *D*'s are contained within the *S*'s. Throughout the list structure a distinction is maintained between programs which do things, the *Z*'s and *P*'s, and those which provide decision-making information, the *D*'s and *Q*'s. The former are analogous to the *O* routines and the latter to the *T* routines in Miller's (1960) TOTE units. Only the lowest level routines can result in the direct execution of a subroutine coded in IPL-V and the higher levels serve only to hold together various combinations of the executable routines. The underlying principle is that the *P*'s and

Q's are the basic information processing capabilities a "subject" possesses, and to perform various tasks one assembles the proper sequence of *P*'s and *Q*'s into *Z*'s and *D*'s. These *Z*'s and *D*'s are then assembled into an *S* which constitutes the simulation program. Ultimately there would exist a wide variety of *P*'s and *Q*'s from which a psychologist would select those he needs to construct the particular set of procedures required for a given type of simulation.

The ability to select *P*'s and *Q*'s from a pool of available routines and then to arrange them into some task related sequence requires that the communication between subroutines be taken care of automatically rather than requiring that one write programs to establish the communication for each unique combination of routines. The solution to this interface problem has been to use a symbol the function of which is solely to possess a description of the inputs and outputs to the routine. Thus the proposed scheme is referred to as a pseudo-code system. Table 2 consists of examples of the second and third level of the list structure. Symbol *P31* is a pseudo-code for routine *P30*, and the description list of *P31* contains an attribute *A1*, the value list of which contains the names of the inputs to *P30*. The attribute *A2* has a value list which consists of the names of the lists in which the outputs of *P30* are to be stored. In the example given, *M11* and *M13* are input symbols and *M1*,

TABLE 2
Typical Z Level and P Level Lists

Symbols			Description
Z1	9-5		
	P31		Vary an attribute valve
	P41	0	Remember search criteria
P31	9-10		
	P30	0	
9-10	0		
	A1		Input
	9-11		
	A2		Output
9-11	9-12	0	
	0		
	M13		
	M11	0	
9-12	0		
	M1		
	M12	0	

M12 are the output lists. At the present time the person assembling the routines into procedures must be sure that the required inputs have been created by the routines preceding any given routine. It does not appear to be too difficult to have a program which will ascertain the existence of inputs before executing a given routine. Thus, the writers have resolved the interface problem by attaching to each routine via a pseudo-code an exact specification of its inputs and outputs which can then be taken care of by automatic rather than manual means.

The advantage of the representation of a simulation program as a list of pseudo-codes is that all of the list processing capabilities of a language such as IPL-V can be applied to the program itself. Two distinct advantages accrue from this approach: one, a program can be written which will interpret the pseudo-code symbols and perform the simulation; second, it provides a vehicle for eventual self-modification of the routine while it is running. The latter objective has not been attempted, but the former is the interpreter which is the central feature of the present system. Because only certain routines on the list structure presented to the interpreter can be executed, a mechanism was required which would enable the interpreter to distinguish between executable and non-executable symbols. The writers adopted the convention that symbols representing described lists were non-executable, whereas symbols representing simple lists were executable. In addition, local symbols were non-executable. Hence, if the interpreter encounters a symbol representing a non-described list, it assumes that the list named by that symbol consists of IPL-V instructions to be executed. If the symbol represents a described list, the interpreter assumes that inputs and outputs exist, and the description list of this symbol is processed. The interpreter uses the value list of the input attribute to stack input symbols in the communication cell *H0* prior to descending one level in the list structure to obtain a new symbol which may or may not be executable. If the symbol at the next lower level is executable, the interpreter turns control over to that routine during its execution. Upon completion of the routine, control returns to the interpreter which takes the stack of output symbols in *H0* and then sends them to the locations named on the value list of the output attribute of the pseudo-code that contained the executable routine.

How would the interpreter process a simulation program such as that specified in Tables 1 and 2? The symbol *S1* would be given to the interpreter; however, *S1* is described, and its description list is empty; hence no inputs will be transmitted. The interpreter descends one level to find a symbol on the *S1* list to execute, but the symbol encountered is *Z1* which is also described. The interpreter then calls upon itself to process *Z1* in the same manner as it did *S1*. The *Z1* list is then entered to find a symbol to execute. The top symbol of *Z1* is *P31* which is also described, but *P31* has symbols on its input attribute value list which are stacked in *H0* by the interpreter. The interpreter descends one level and encounters the symbol *P30* which is not described; hence the routine *P30* is executed. Upon completion the subroutine *P30* leaves its outputs in *H0* and control returns to the interpreter which ascends one level and uses the symbols on the output attribute value list of *P31* to store the symbols left in *H0* by the *P30* routine. The interpreter then obtains the next symbol on the *Z1* list and repeats the same sequence. When the end of the *Z1* list is reached, the interpreter ascends one level and the value list of the output attribute is inspected to ascertain whether outputs are to be processed. The interpreter then proceeds to the next symbol on the *S1* list, and the whole process is repeated again until a zero pseudo-code is reached on the highest level list which causes the routine to terminate. Fundamentally, the interpreter is an ordinary IPL-V recursive program which calls upon itself to work its way up and down the branches of the list structure representing the simulation program. A key feature of the recursive property of the interpreter is that it maintains its own current instruction address list which is pushed down and popped up as the list structure is processed.

One additional feature is required in order to obtain adequate flexibility in the logic of the simulation program. The test cell of IPL-V (*H5*) which can have the values $+$ or $-$, is normally used to effect bi-directional branching. Local symbols have been reserved to represent the location to which the program will branch. When the interpreter encounters a local symbol on a list at any level, it checks the value of *H5*. If *H5* is positive, it will obtain the symbol following the local symbol and continue processing. If *H5* is negative, the interpreter goes to the location named by the local symbol to obtain the next symbol processed. The value of *H5* can be set

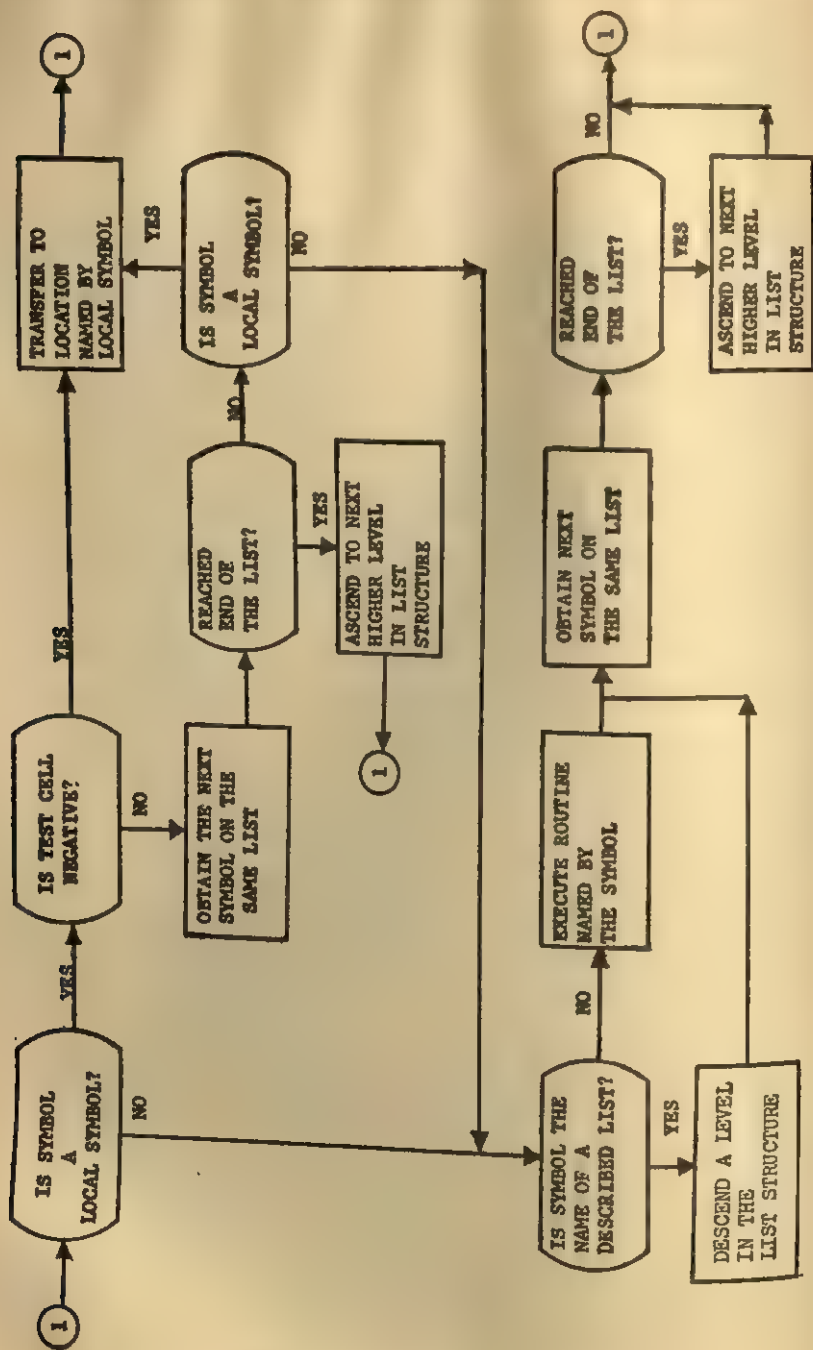


Figure 1. Interpreter Flow Chart.

only by the decision-type routines, namely *Q*'s and *D*'s. Thus, the local symbol provides a convenient way to modify the flow of processing within the program as a function of the program itself. Figure 1 presents a flow chart of the interpreter as described above.

Summary

The representation of a simulation program as a list of symbols which are to be executed is a common practice in IPL-V programming, as it allows one to treat a program as if it were information to be processed. The use of such lists to contain a pseudo-code, the symbols of which represent executable or non-executable routines, constitute a departure from usual practice. The departure represented by a pseudo-code system and its associated recursive interpreter is one in the direction of increased sophistication and operational flexibility. Because much of the power of a simulation program written under this system rests upon the executable routines, one is forced continually to search the behavior being simulated for fundamental information processing modules which appear in many contexts. The pseudo-code scheme permits one to use these basic modules in a wide variety of contexts without writing the unique instructions to handle the between routine communication of information. Although this scheme was developed within the context of a particular simulation project, the writers feel that the general approach could profitably be used in many other areas of simulation.

REFERENCES

- Baker, F. B. "An IPL-V Program for Concept Attainment." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIV (1964), 119-127.
- Miller, G. A., Galanter, E., and Pribam, K. *Plans and the Structure of Behavior*. New York: Holt, 1960.
- Newell, A., et al., (Editors). *Information Processing Language-V Manual, Second Edition*. Englewood Cliffs: Prentice Hall, 1964.



A FORTRAN GENERATOR OF POLYNOMIALS ORTHONORMAL OVER UNEQUALLY SPACED AND WEIGHTED ABSCISSAS

PHILLIP L. EMERSON
Washington State University

THE majority of the applications of orthogonal polynomials for the analysis of the trend of a set of data into components of different polynomial degrees, has been confined to experimental data from the laboratory, since the orthogonal polynomials traditionally used are applicable only when the levels of the independent variable are equally spaced and when equal numbers of observations have been made at the different levels. Certain other special cases are known, however, such as the Krawtchouk polynomials which are orthogonal over equally spaced abscissas, under a set of weights proportional to the binomial coefficients, and the Poisson-Charlier polynomials which are similar but orthogonal under weights proportional to the ordinates of the Poisson distribution (Szegő, 1959). Such special sets can be useful when the numbers of observations at the levels of the independent variable are proportional to the weights, since C_j , the coefficient of $P_j(x)$, the orthogonal polynomial of degree $j - 1$ in the regression equation

$$Y_x = C_1 P_1(x) + C_2 P_2(x) + \dots + C_n P_n(x) \quad (1)$$

is given by

$$C_j = \frac{\sum_{i=1}^n w_i P_j(x_i) Y_i}{\sum_{i=1}^n w_i P_j^2(x_i)} \quad (2)$$

and SS_j , the independent portion of the between- x sum of squares used to test the contribution of $P_j(x)$ for statistical significance, is given by

$$SS_j = \frac{r \left[\sum_{i=1}^n w_i \bar{Y}_i P_j(x_i) \right]^2}{\sum_{i=1}^n w_i P_j^2(x_i)} \text{ with 1 df} \quad (3)$$

where there are rw_i observations of Y at the point x_i , and \bar{Y}_i is the mean of these (Emerson, 1965). The condition of orthogonality necessary here is that $\sum_{i=1}^n w_i P_j(x_i) P_k(x_i) = 0$, when $j \neq k$. If, in addition, $\sum_{i=1}^n w_i P_j^2(x_i) = 1$ for $j = 1, 2, \dots, m$, the polynomials are said to be orthonormal and equations (2) and (3) are somewhat simpler.

Grant (1956) has outlined methods of analysis covering a number of designs likely to occur in educational and psychological research where there may be between-subjects variables, within-subjects variables, or both. Although his description involved only the special case of equal spacing and weighting of the levels of the trend variable, the general method of setting up the designs of higher complexity are readily generalizeable for unequal spacing and weighting.

The FORTRAN program described below generates a set of orthonormal polynomials for any input set of n ordered abscissas x_i ($x_1 < x_2 < \dots < x_n$) and n positive weights w_i for $i = 1, 2, \dots, n$. It will thus provide sets of polynomials which can be used in any of the analyses covered by Grant, regardless of the spacing and weighting of the levels of x .

For maximum versatility, it is written as a subroutine so that any option may be elected re. the formats of input and output to and from the calling program which the user must provide. The method involves the "orthogonalization" (attributed to Stone, 1928, by Szegő, 1959) of the m functions $1, x, x^2, \dots, x^{m-1}$, and is essentially the same as the Gram-Schmidt process (Paige and Swift, 1961, p. 78) and the method described by Robson (1959) and Wiener (1949, p. 32), except that here w_i need not be constant for $i = 1, 2, \dots, n$.

To obtain a set of m polynomials $P_1(x), P_2(x), \dots, P_m(x)$, orthonormal over the n ordered abscissas x_i ($x_1 < x_2 < \dots < x_n$), under the n positive weights w_i for $i = 1, 2, \dots, n$ where $m \leq n$, the orthogonalization is accomplished by the recursive process

$$P_1(x) = \frac{1}{\sqrt{\sum_{i=1}^n w_i}}$$

$$Q_j(x) = x^{j-1} - \sum_{i=1}^{j-1} P_i(x) \sum_{i=1}^n w_i x_i^{j-1} P_i(x_i),$$

in which the normalization

$$P_j(x) = \frac{Q_j(x)}{\sqrt{\sum_{i=1}^n w_i Q_j^2(x_i)}}$$

is performed at each step for $j = 2, 3, \dots, m$.

The procedure is best understood by noting that a least-square linear combination of the $j - 1$ functions $P_v(x)$ for $v = 1, 2, \dots, j - 1$ is fitted to the n values x_i^{j-1} for $i = 1, 2, \dots, n$, and subtracted from x^{j-1} . The residual polynomial $Q_j(x)$ must be of degree $j - 1$ and orthogonal to all $P_v(x)$ for $v \leq j - 1$, which are of lower degree.

In order to keep the exponentiated numbers within a manageable range, the program first transforms x to z , the standard deviate of the x distribution, i.e.,

$$z = \frac{x - \frac{\sum w_i x_i}{\sum w_i}}{\sqrt{\frac{\sum w_i x_i^2}{\sum w_i} - \left(\frac{\sum w_i x_i}{\sum w_i}\right)^2}} \quad (4)$$

where all summations are from $i = 1$ to n , and constructs the polynomials with the argument z .

The program, called SUBROUTINE ORNOPL, takes as inputs m , the number of orthonormal polynomials to be constructed, the highest degree of which is $m - 1$; n , the number of abscissas; x_i , the values of the abscissas; and w_i , the weights or numbers of observations at the respective abscissas. It determines the values of the orthonormal polynomials in z at z_i , $i = 1, 2, \dots, n$, and stores them in the $m \times n$ matrix $P_{j,i}$, $j = 1, 2, \dots, m$, $i = 1, 2, \dots, n$. It also determines the values of the coefficients of the powers of z in each orthonormal polynomial, and these are denoted $C_{j,k}$ for $j = 1, 2, \dots, m$, $k = 1, 2, \dots, j$, where

$$P_1(z) = C_{1,1}$$

$$P_2(z) = C_{2,1} + C_{2,2}z$$

$$P_3(z) = C_{3,1} + C_{3,2}z + C_{3,3}z^2,$$

etc.

In the first row of the C matrix, since only the first cell is occupied, the rest of the row is used for working storage.

The outputs of SUBROUTINE *ORNOPL* to the calling program are the $P_{j,i}$ matrix; the $C_{j,i}$ matrix; z_i for $i = 1, 2, \dots, n$; $XBAR$ and SDX which are respectively the subtracted term in the numerator and the radical of the denominator of the right-hand side of equation (4). All inputs are left unchanged.

If the regression equation is to be used for interpolation of values of Y , or if pairs of interpolated values of x and $P_j(z)$ are desired, the values of $XBAR$ and SDX can be used to first transform x to z which is then substituted into the appropriate equation.

Testing of the program indicates that reasonable accuracy may be expected with single precision arithmetic for m up to about 6 or 7. With double precision arithmetic this is extended to about 10. A simple accuracy check can be made by examining the $P_{j,i}$ matrix. There should be $j - 1$ sign changes in the j th row, and if $m = n$, there should be $n - i$ sign changes in the i th column.

Subroutine *ORNOPL* has been timed on the IBM 709 at Washington State University for m up to 15 and n up to 50, where n is the number of abscissas and m is the number of polynomials in the generated set. A good approximation of the amount of time consumed each time it was called, is given by the equation $t = .001n(m - 1)^2$, where t is in seconds.

APPENDIX

SUBROUTINE ORNOPL($M, N, X, W, P, C, XBAR, SDX, Z$)
PROGRAMMED IN FORTRAN II DEC. 1964 BY PHILLIP

```

C      EMERSON
D      DIMENSION X(100), W(100), P(20, 100), C(20, 20), Z(100)
D 100  SUM = 0.0
D 110  SUMX = 0.0
D 120  SUMSQ = 0.0
D 130  DO 160 I = 1, N
D 140  SUM = SUM + W(I)
D 150  SUMX = SUMX + X(I)*W(I)
D 160  SUMSQ = SUMSQ + X(I)**2*W(I)
D 170  XBAR = SUMX/SUM
D 180  SDX = SQRTF(SUMSQ/SUM - XBAR**2)
D 190  DO 200 I = 1, N
D 200  Z(I) = (X(I) - XBAR)/SDX
D 210  C(1, 1) = 1.0/SQRTF(SUM)
D 220  DO 230 I = 1, N
D 230  P(1, I) = C(1, 1)
D 240  DO 460 K = 2, M
D 250  KM1 = K - 1

```

```

260 DO 290 J = 2, K
D 270 C(1, J) = 0.0
280 DO 290 I = 1, N
D 290 C(1, J) = C(1, J) + W(I)*Z(I)**KM1*P(J - 1, I)
300 DO 330 I = 1, N
D 310 P(K, I) = Z(I)**KM1
320 DO 330 J = 2, K
D 330 P(K, I) = P(K, I) - C(1, J)*P(J - 1, I)
D 340 SUM = 0.0
350 DO 360 I = 1, N
D 360 SUM = SUM + W(I)*P(K, I)**2
D 370 C(K, K) = 1.0/SQRTF(SUM)
380 DO 390 J = 2, K
D 390 C(1, J) = C(1, J)*C(K, K)
400 DO 440 J = 2, K
410 JM1 = J - 1
D 420 C(K, JM1) = 0.0
430 DO 440 I = J, K
D 440 C(K, JM1) = C(K, JM1) - C(1, I)*C(I - 1, JM1)
450 DO 460 I = 1, N
D 460 P(K, I) = P(K, I)*C(K, K)
470 RETURN
END

```

REFERENCES

- Emerson, P. L. "Orthogonal Polynomials for Unequally Weighted Means." *Biometrics*, XXI (1965), (in press).
- Grant, D. A. "Analysis-of-Variance Tests in the Analysis and Comparison of Curves." *Psychological Bulletin*, LIII (1956), 141-154.
- Paige, L. J. and Swift, J. D. *Elements of Linear Algebra*. Los Angeles: Ginn and Company, 1961.
- Robson, D. S. "A Simple Method of Constructing Orthogonal Polynomials when the Independent Variable is Unequally Spaced." *Biometrics*, XV (1959), 187-191.
- Stone, M. H. "Developments in Hermite Polynomials." *Annals of Mathematics (Series 2)*, XXIX (1928), 1-13.
- Szegő, G. *Orthogonal Polynomials*. New York: American Mathematical Society Colloquium Publications, 1959.
- Wiener, N. *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Cambridge, Mass.: The M.I.T. Press, 1949.



A SUBROUTINE TO REFINER THE INVERSE OF A MATRIX

NATHAN JASPEN
New York University

THE calculation of the inverse of a matrix involves many multiplications and divisions, often with the result of a substantial loss of significant figures in the elements of the inverse. In recognition of this, Von Neumann (1959) advised that ten or twelve significant digits be retained in computer operations, even though a smaller number in the final result were sufficient for most practical purposes. It sometimes happens, in matrix inversion and in Doolittle procedures, that only one or two significant figures in the result are correct, although a much larger number of digits are retained in all calculations (Von Neumann and Goldstine, 1947).

Hotelling (1943) provided an efficient iterative technique to refine the inverse of a matrix, given a reasonably good approximate inverse. If a matrix A and its approximate inverse C_0 are given, then a set of improved inverses C_1, C_2, \dots , may be calculated using the formula

$$C_{i+1} = C_i(2 - AC_i) \quad (1)$$

where the matrix 2 is a square matrix of the same order as A , with twos in the main diagonal and zeros elsewhere. Householder (1953) refers to (1) as the Hotelling-Bodewig iteration formula.

Ullman (1944) found that it was possible to hasten convergence by using the formula

$$C_{i+1} = C_i[1 + (1 - AC_i) + (1 - AC_i)^2 + (1 - AC_i)^3 + \dots]. \quad (2)$$

However, the main problem is that the iterative process, using either formula, may be divergent, rather than convergent. In this respect it is unlike the Newton procedure for square root, which

necessarily converges. The conditions for convergence of (1) are given by Hotelling (1943). No matrix inversion procedure that the writer has ever used has resulted in an inverse so poor that (1) diverged; however, an inverse with gross discrepancies would almost certainly diverge.

In the following Fortran subroutine, based on Hotelling's formula (1), matrices A and C , are available on entrance, and the final inverse C is available on exit. The number of iterations (IT), and the sum of the squares of the deviations of the elements of the product matrix AC from the elements of the identity matrix ($ERSQ$) are also available on exit. Ordinarily, only a small number of iterations is required, since the converging process accelerates. The number of decimal places of sure accuracy increases in geometric progression from iteration to iteration. Hotelling notes that the iterative method "will always be the most efficient if a sufficiently large number of decimal places is required" (1943, p. 15).

The iterations continue as long as the deviate sum of squares is reducing. When the process is divergent, or when a stable inverse is attained, the iteration procedure terminates. This test tends to lengthen the subroutine somewhat; however, it does permit greater accuracy in the result than does the adoption of some arbitrary criterion of precision.

The dimension statement may be altered to meet the size of the computer. For high accuracy, the subroutine should be written in the double-precision mode.

```
SUBROUTINE INV (A, C, M, IT, ERSQ)
  DIMENSION A(20, 20), C(20, 20), P(20, 20), D(20, 20)
  COMMON A, C, M, IT, ERSQ
```

```
100  IT = 0
      ERSQ = 0
      DO 110 I = 1, M
        DO 110 J = 1, M
          P(I, J) = 0
          DO 110 K = 1, M
110    P(I, J) = P(I, J) + A(I, K) * C(K, J)
          DO 120 I = 1, M
120    P(I, I) = P(I, I) - 1.0
```

```

DO 130 I = 1, M
DO 130 J = 1, M
130 ERSQ = ERSQ + (P(I, J)) ** 2
140 DO 150 I = 1, M
150 P(I, I) = P(I, I) - 1.0
DO 160 I = 1, M
DO 160 J = 1, M
160 P(I, J) = -P(I, J)
DO 170 I = 1, M
DO 170 J = 1, M
D(I, J) = 0
DO 170 K = 1, M
170 D(I, J) = D(I, J) + C(I, K) * P(K, J)
DO 180 I = 1, M
DO 180 J = 1, M
P(I, J) = 0
DO 180 K = 1, M
180 P(I, J) = P(I, J) + A(I, K) * D(K, J)
DO 190 I = 1, M
190 P(I, I) = P(I, I) - 1.0
F = 0
DO 200 I = 1, M
DO 200 J = 1, M

200 F = F + (P(I, J)) ** 2
IF (ERSQ - F) 230, 230, 210
210 ERSQ = F
IT = IT + 1
DO 220 I = 1, M
DO 220 J = 1, M
220 C(I, J) = D(I, J)
GO TO 140
230 RETURN
END

```

REFERENCES

- Hotelling, H. "Some New Methods in Matrix Calculation." *Annals of Mathematical Statistics*, XIV (1943), 1-34.
- Householder, A. S. *Principles of Numerical Analysis*. New York: McGraw-Hill, 1953.
- Ullman, J. "The Probability of Convergence of an Iterative Process

of Inverting a Matrix." *Annals of Mathematical Statistics*, XV (1944), 205-213.

Von Neumann, J. *The Computer and the Brain*. New Haven: Yale University Press, 1959, 24-28.

Von Neumann, J. and Goldstine, H. H. "Numerical Inverting of Matrices of High Order." *Bulletin of the American Mathematical Society*, LIII (1947), 1021-1099.

THE CALCULATION OF PROBABILITIES CORRESPONDING TO VALUES OF z , t , F , AND CHI-SQUARE

NATHAN JASPEN
New York University

THE F ratio may be normalized by means of a transformation (Kelley, 1947; Kendall, 1955), as follows:

$$z = \frac{\left(1 - \frac{2}{9j}\right)F^{1/3} - \left(1 - \frac{2}{9i}\right)}{\left(\frac{2}{9j}F^{2/3} + \frac{2}{9i}\right)^{1/3}} \quad (1)$$

where F is the ratio of two independent variances with i and j degrees of freedom respectively, provided that F is not less than 1. The variable z is normally distributed with zero mean and unit variance.

Wilson and Hilferty (1931) found that $(\chi^2/n)^{1/2}$ is nearly normally distributed with mean $1 - 2/9n$ and variance $2/9n$, for values on n greater than 3. The fit is better than R. A. Fisher's transformation $(2\chi^2)^{1/3} - (2n - 1)^{1/2}$, which is normally distributed with unit variance for large n . The transformation of F to z is due to Paulson (1942), who expressed F in the form of two χ^2 's, each divided by its degrees of freedom. He used the results of Wilson and Hilferty (1931), and also some work by Fieller (1932) on the distribution of the ratio of two normally distributed variates.

According to Kelley (1947, 1948), this transformation is close if $j > 3$. If $j < 3$, he recommends that the following value be substituted:

$$z' = z(1 + .0800z^4/j^3). \quad (2)$$

The case for $j = 3$ is left open. The writer found empirically that the correction formula (2) yields substantially better results for $j = 3$ than does the application of (1) alone. Incidentally, the correction formula in Kelley (1947) suffers from a misprint.

Given the value of z , corrected if necessary, the value of the corresponding probability p is available from the following approximation (Zelen and Severs, 1964):

$$p = .5/(1 + c_1z + c_2z^2 + c_3z^3 + c_4z^4)^4, \quad (3)$$

where

$$\begin{aligned} c_1 &= .196854, \\ c_2 &= .115194, \\ c_3 &= .000344, \\ c_4 &= .019527. \end{aligned} \quad (4)$$

This approximation is based on work by Hastings (1955). He has also made available closer approximations, involving equations of higher order. Smillie and Anstey (1964) present another version of (3). As far as the writer knows, Smillie and Anstey were the first to connect the F and z transformation with a Hastings approximation.

If $F < 1$, find the value of p corresponding to the reciprocal of F , with the values i and j interchanged, and then subtract the p so found from unity. The absolute value of the numerator of (1) should be used.

To transform Student's t , use the familiar fact that

$$F = t^2, \quad (5)$$

and let $i = 1$, and set j equal to the degrees of freedom corresponding to t .

To transform χ^2 , use the fact that

$$F = \frac{\chi^2}{\text{degrees of freedom}}, \quad (6)$$

and that $j = \infty$. It follows that

$$z = \frac{F^{1/3} - (1 - 2/9i)}{(2/9i)^{1/3}}. \quad (7)$$

If $F < 1$, substitute the reciprocal of F into (7), and subtract the resulting p from unity.

Table 1 shows how the computed values of p compare with the tabled values of p for the most ill-fitting combination of i and j tested, namely, for 2 and 2 degrees of freedom, for selected values of F .

A Fortran subroutine that can be used to find p , given F , i , and j , is appended.

TABLE 1

A Comparison of Computed and Tabled Values of p for Selected Values of F with 2 and 2 Degrees of Freedom

F	Computed p	Tabled p
999.0	.0010	.001
199.0	.0040	.005
99.0	.0083	.010
39.0	.0229	.025
19.0	.0489	.050
9.0	.1007	.100
3.0	.2503	.250
1.0	.5000	.500
.3333	.7498	.750
.1111	.8993	.900
.0526	.9512	.950
.0256	.9772	.975
.0101	.9917	.990
.0050	.9960	.995
.0010	.9990	.999

SUBROUTINE FTOP (F , I , J , P)

$P = 1.0$

IF (F) 100, 100, 10

10 IF (I) 100, 100, 20

20 IF (J) 100, 100, 30

30 IF ($F - 1.$) 40, 50, 50

40 $B = J$

$W = I$

$G = 1./F$

GO TO 60

50 $B = I$

$W = J$

$G = F$

60 $ALPHA = 2./(9.*B)$

$BETA = 2./(9.*W)$

$TOP = (1. - BETA) * G ** (1./3.) - 1. + ALPHA$

```

      BOT = SQRTF (BETA * G ** (2./3.) + ALPHA)
      Z = ABSF (TOP/BOT)
      IF (W - 3.) 70, 70, 80
70  Z = Z * (1. + .0800 * Z ** 4/W ** 3)
80  CA = .196854
      CB = .115194
      CC = .000344
      CD = .019527
      P = .5/(1. + Z * (CA + Z * (CB + Z * (CC + Z * CD)))) ** 4
      IF (F - 1.) 90, 100, 100
90  P = 1. - P
100 RETURN
      END

```

REFERENCES

- Fieller, E. C. "The Distribution of the Index in a Normal Bivariate Population." *Biometrika*, XXIV (1932), 428-440.
- Hastings, C., Jr. *Approximations for Digital Computers*. Princeton: Princeton University Press, 1955.
- Kelley, T. L. *Fundamentals of Statistics*. Cambridge: Harvard University Press, 1947, 325-331.
- Kelley, T. L. *The Kelley Statistical Tables* (Revised). Cambridge: Harvard University Press, 1948.
- Kendall, M. G. *The Advanced Theory of Statistics*, Vol. II. London: Charles Griffen, 1955, 116-118.
- Paulson, E. "An Approximate Normalization of the Analysis of Variance Distribution." *Annals of Mathematical Statistics*, XIII (1942), 233-235.
- Smillie, K. W. and Anstey, T. H. "A Note on the Calculation of Probabilities in an F-Distribution." *Communications of the ACM* (Association for Computing Machinery), VII (1964), 725.
- Wilson, E. B. and Hilferty, M. M. "The Distribution of Chi-Square." *Proceedings of the National Academy of Sciences*, XVII (1931), 684-688.
- Zelen, M. and Severs, N. C. "Probability Functions." In *Handbook of Mathematical Functions* (National Bureau of Standards Applied Mathematics Series No. 55). Washington: U. S. Government Printing Office, 1964, p. 932.

SCORIT—A FORTRAN PROGRAM FOR SCORING AND ITEM ANALYSIS OF PORTA-PUNCH TEST CARDS¹

PAUL A. GAMES

Ohio University

ALTHOUGH large institutions may have equipment that converts from mark-sense or optical input to computer-acceptable punched cards, many small computer installations are faced with the problem of manually punching cards from test records. This costly and error-introducing procedure can be avoided by having each student punch his own machine-acceptable test record through using porta-punch cards as an answer sheet. The card shown in Figure 1 is employed for data input in a program used on the IBM 1620 with 40K memory and disk, but this program could be readily modified for use on other computers with auxiliary storage devices.

Input

An identification card is followed by the correct answer card (key) for the test. The number of questions (1-70) is punched as the seat number on the porta-punch card, and the correct answer (A to E or *omit*) is hand punched to complete the key. If an item is left blank on this key card, then OMIT is counted as the correct response. Any person also omitting that item will gain one point, while any person punching it does not. This permits the use of six alternatives in a question by listing the sixth response as "F"—(leave blank). Any items with from 2-6 responses may thus be used, in any sequence or mixture desired. The key card is followed by the student test cards (up to 692 students). Each student card should have a seat number and must have no double punches in any

¹ Publication of this report was made possible through a grant from the University Research Committee of Ohio University.

SEAT NUMBER	NAME	FIRST NAME										COURSE										TEST										CARD NO. OR TEST NO.								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30		31	32	33	34	35			
000		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	
111		B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
222		C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
333		D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	
444		E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	
555		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	
666		B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	
777		C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
888		D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	
999		E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	

NEW D61014

DATE

MISC.

Fig. 1. Porta-punch card D61014, the data card for SCORIT.

column. The end of the deck is indicated by a card with a seat number of 999.

Output

The cards are scored in the order in which they are fed to the computer, and the seat number and score are punched on cards. By sorting by seat number, this output may be used to produce copies for posting in class or for records, or may be used as inputs to further programs for keeping the course records. After the last card is read, the frequency distribution of scores, mean, standard deviation, crude median, and number of scores are punched out.

The final output is the item analysis for each item used, reported in the form shown below, with the phi coefficient computed on the upper-half versus lower-half of the scores as one dimension, and pass-fail as the other. The output is designed so that it may be cut and pasted crosswise on the back of a 4×6 or larger card used for item file storage.

		OMIT	A	B	C	D	E	
QUESTION 1	{	SAMPL	0	1	1	9	6	8 U
QUESTION 1	{	DATA	0	6	0	4	15	0 L E (correct answer)
		PHI -	.43, PERCENT PASS - 16.					

QUESTION 2	{	SAMPL	0	12	6	2	2	3 U
QUESTION 2	{	DATA	0	9	3	6	2	5 L A (correct answer)
		PHI -	.12, PERCENT PASS - 42.					

The Kuder-Richardson formula 20 reliability coefficient completes the output.

Usage of the Porta-punch Test Cards²

The cards are placed on a smooth surface while a student is taking the test. The student is instructed to circle his seat number and the correct answer to each question. If he wishes to change an answer, he merely crosses out the previous answer and circles a new answer. Upon completing the test, he places the answer card upon a sponge, and using a plastic stylus, he punches out his seat number and the circled answers. The styli and sponges may be secured

² SCORIT and a more complete set of operative procedures, and instructions to the students may be obtained through the 1620 Users Group Library. SCOR2C, a modification that permits the use of two cards per student (and hence up to 140 questions) may be obtained from the author.

from representatives of Burgess Cellulose Co. (Freeport, Ill., 61033) at nominal cost. These and the answer cards are typically handed out by an assistant at the beginning of the testing period. Thus only the tests themselves are left to be distributed with care. One minute before the end of the test period, the tester should give a warning to finish the test and to start punching. The only student resistance to the use of these cards has been from slow-reading students who have not finished the test, and who would prefer to spend this minute on unattempted items. The collection of materials at the end of the test is slightly more trouble than it is with the usual separate electrographic answer sheet, since four items are to be collected rather than three.

Students who follow the instructions should have no double punches. If a double punch (or very careless punching) is encountered, the card reader will reject the card, and will sort it into the reject pocket. This card may be removed and the card reader started again; the program will be unaffected. Any such rejected cards may be rapidly hand scored by simply placing the test card on top of the answer card, by counting the number of coinciding holes, and by checking any OMIT columns. Rejected cards will not be included in the computer output. The author has adopted the rule that a test card rejected by the reader costs the student one test point, and each double punch costs an additional point. This improves the quality of card punching rapidly, but it is still wise to have spare cards on hand for the over-confident student who ignores the directions and attempts to punch each answer as he goes through the test.

DATA PROCESSING PROCEDURES TO IMPROVE CLASSROOM TESTING

QUENTIN C. STODOLA
Northern Michigan University

To what extent can the use of commonly available data processing equipment improve practices in classroom testing? A partial answer to this question was given in procedures developed in an experiment conducted in a college course in tests and measurements at North Dakota State University. This experiment was supported by the Cooperative Research Branch, U. S. Office of Education.¹

There were two purposes to the experiment: one was to determine the effect of frequent testing on learning; the other was to demonstrate how the use of data processing methods may facilitate classroom testing. Results of the experiment indicated that frequent testing increased learning. A detailed description of the research findings is available in the project report which may be obtained from the author.

In the experiment, data processing methods were used to facilitate test construction in the following manner: (1) over a thousand pretested items were punched on IBM cards; (2) these items were assigned a number code based on a detailed content analysis of course objectives as well as on the statistical characteristics of the items; (3) through use of the code system, tests were constructed by selecting items possessing the specific content and statistical characteristics desired; and (4) after the items in punch card form were selected by use of the number code system, they were pro-

¹ A detailed description of this experiment may be found in the following monograph: Stodola, Quentin C., Eustice, D. Edward, Kolstoe, Ralph H. *Frequent Classroom Testing as a Learning Aid Using Data Processing*, North Dakota State University, 1964.

cessed on an IBM 407 Accounting Machine to cut mimeograph stencils which were in turn used for convenient reproduction of test booklets. Thus it was possible to construct and to reproduce rapidly a series of classroom tests which were tailor-made to given specifications.

Student responses to test items were recorded on mark-sense cards and were analyzed through using an SPS 1620 IBM program. Results of the item analysis were examined to identify those items which the class found difficult. The class then engaged in discussion of the points covered by these difficult items. This discussion was followed by further testing on the same points. The experimenters believe that the success of the frequent testing was dependent on the class discussion—that learning was improved through further consideration of those points identified by item analysis as being difficult.

In addition to the procedures developed in the experiment, the author has recently taken another step to increase the value of the IBM card item file. He has prepared a textbook in tests and measurements, in which numbers are printed on the page margins. These numbers, which are consistent with the test item code system, enable a student who misses a particular item to note the code number of that item printed in the test booklet and to refer to the section of the textbook which relates to the missed item.

The data processing procedures described provide a convenient approach to the process of teaching, testing, re-teaching, and re-testing. Experience with this approach indicates that it is an effective aid to student learning.

VARIABLES AFFECTING THE GRADUATE ASSISTANT IN A COMPUTER TRAINING POSITION¹

M. GORDON HOWAT

Cornell Aeronautical Laboratory, Inc.
of Cornell University

THE aim of this experiment was to evaluate those variables contributing to the maximum learning and productivity in training a programmer by having him serve as an assistant in a computer center. This experiment, using an intensive questionnaire, was profitable even if only applied to the same computer center the following year. The lack of literature specifying major variables in various procedures of programmer training begs for good research, and a wise choice among alternative procedures depends upon it.

Procedure

Selection and Setting. In early September, professors suggested to graduate students with an appropriate mathematics or statistical background that if they had any interest in an assistantship in the Computer Center they should interview the Director. The first six showing an initial interest were accepted.

These graduate students, hereafter called *Ss*, began a programming and operating career in a small university setting, with a new IBM 1620-40K installation and with a Director (*D*), who had previously served in a social and behavioral science section at a large university. The goal and task of this team was to meet all computational needs of campus research, of regional NIH-Supported research, and of the state mental health institutions.

¹This study was conducted at New Mexico Highlands University. The analysis of results was accomplished by the author at Cornell Aeronautical Laboratory. The critical reading of H. D. Sherrerd is appreciated.

Assistants (Ss) and Instruction. All Ss were 22 to 26 years of age. Four, S_1 to S_4 , were mathematics majors, S_5 , a physics major, and S_6 , a psychology major. Only S_3 had had some previous problem and operating experience on the IBM 1620. The foreign student, S_5 , had had formal FORTRAN I instruction. Two had previously spent a quarter at the institution. All except S_4 , who began in the following Winter (W) quarter, attended 18 hours of formal FORTRAN II instruction, during evening and Saturday morning hours in early October.

Assistantship Policy and Conditions. In his first weekly meeting D urged that cooperation should be used wherever possible. This statement was the only mention of cooperation. It was also requested at this time that all administrative problems be discussed only among this group, defined as the "computer staff" (an in-group), rather than outside. Each S was issued a key to the computer room to enable private evening as well as semiprivate daily use of the computer.

For each quarter, D posted his own weekly schedule. The proportion of the work day D had available from other duties to guide Ss and to answer questions about computation was as follows: Fall (F) 0.95, Winter (W) 0.5, and Spring (Sp) 0.85. In W and Sp quarters tasks were listed in the job-book for the Computer Room by date prior to direct discussion of the task.

Evaluation. A psychologist (also D) designed and administered two questionnaires to assess the major variables. All items were carefully ordered so that responses would not be biased by the contents of preceding questions. In mid-June, each S was unexpectedly given a six-item multiple-choice questionnaire which included rankings of quarters. A 14-item structured interview with D followed immediately. Independently, D used his own rating of Ss' achievement based on trained personnel as the standard.

All P values are exact probabilities for the response pattern to the item based upon multinomial expansion of the number of alternatives and the number choosing. Seven items are not reported. One item lacked clarity, and six were introductory or involved personal goals.

Results

The Ss required a mean of 3.7 mos. practice before they were

capable of programming in FORTRAN all user needs without help from *D*, although *D*'s comments could have led to greater, but unneeded, efficiencies in running time. The standard deviation (*S.D.*) was 2.7 mos., and individual scores were as follows: S_1 —2.0, S_2 —4.8, S_3 —5.5, S_4 —1.0, S_5 —1.0, and S_6 —8.0 mos. These times appeared related to capability; independent of whether the *S* served 5 or 10 hrs. per week.² The practice required for *S*s to become proficient computer operators were the following: S_1 —0.3, S_2 —0.3, S_3 —0.2, S_4 —0.4, S_5 —2.2, and S_6 —0.7 mos. The mean time was 0.7 mos., with $S.D. = 0.2$ mos. The range in *S*s' individual overall accomplishment at typical assistantship wages was such as to require some 34 to 19 hrs. per week for the lowest to highest achieving *S*, respectively, to match the value of service from fully trained programmers and operators, according to *D*'s comparative assessment. The individual's assessment by quarter has been presented elsewhere (Howat, 1964).

The mean overall accomplishment per quarter, as rated by *D*, was $F = 5.2$, coefficient of variation (*C.V.*) = 0.28; $W = 7.3$, *C.V.* = 0.10; $Sp = 8.0$, *C.V.* = 0.25, where 10 is the maximum based upon a trained programmer or operator. All important variables, viz., *S*'s overall ability, interstaff rapport, and amount of responsibility delegated by *D*, increased with time.

The *S*'s ranking of quarters disagreed with an orderly time trend only in three items. The *W* quarter was ranked lowest, *F* second, and *Sp* the highest for the quarter in which: (1) manner of job assignment, written or oral (a dummy item³) was most closely achieved ($P = .008$), (2) communication between *S* and *D* was most satisfactory and responsive ($P = .02$), and (3) *S*s' feeling of accomplishment. The feeling of accomplishment was poorest in the *W* quarter ($P = .02$) rather than in the *F* quarter, while it was highest in the *Sp* quarter ($P = .08$). In relation to (3) above, *S*s' reasons for this ranking can be summarized as follows: *F*—much was learned ($P = .04$); *W*—the least was accomplished or, for two *S*s accomplishment was intermediate ($P = .08$); *Sp*—greater re-

² The average time served per week across quarters due to partial teaching assignments for some *S*s were $F = 8.0$, $W = 6.0$, and $Sp = 6.0$ hrs. Of this time the approximate mean spent in non-programming activities, e.g., operating, maintenance, and assisting was as follows: $F = 44\%$, $W = 36\%$ and $Sp = 28\%$.

³ This dummy variable, allowing the projection of a non-real difference to quarters, served the purpose of allowing indirect expression of an attitude to quarters.

sponsibility and progress, a gain in interest for S_3 , and for S_1 and S_6 , distracting study pressures ($P = .08$). The S s confirmed that they had learned the most in the F quarter ($P = .004$).

The greatest improvement S s would have desired in each quarter centered around the following: F —the newness of the center, e.g., keeping the users out of the Computer Room, ($P = .01$); W —more stimulating problem-solving conditions, e.g., bigger time blocks, more difficult problems, and consulting directly with user ($P = .04$); Sp —predominately the same stress upon an efficient problem-solving environment ($P = .01$).

All S s described the computer center atmosphere as cooperative and indicated that the amount of informal cooperation they had experienced was their preference⁴ ($P = .016$).

In response to the question, "What corresponded with your greatest happiness or enthusiasm in the work?" most requested original problems to diagnose and to program, and the rest named learning to program ($P = .02$). To the follow-up question "Did output?", answers were equally divided between a "Yes" response and alternatively naming freedom of approach or their own appraisal as more important. Likewise, having appropriate problems to solve was the factor which contributed most to making S 's computer center service enjoyable ($P = .008$) and study of computer-related items the most profitable ($P = .08$). To have computational skills as a tool for future problems was another motivation offered by two S s.

All S s felt that the best use of a graduate assistant in a computer center should be to program and to add any suggestions considered appropriate to problems in their area of specialization ($P = .001$). Also, two expressed the desire to serve as consultants or advisors to students or beginning programmers.

Discussion and Conclusions

The criteria of selection were found adequate to achieve homogeneous and high interest in the programming instruction and in the following working experience. All S s enjoyed their identification as a group, and the overall rapport increased with time. Yet, the range among S s' achievement was two to one. Motivation was high, and it contributed only a small amount to this range in achievement.

⁴ The item answered was, "Would more competition or more cooperation with other programmers have achieved better results?"

The highest achiever, S_4 , obtained no formal instruction. Thus, it is concluded that the most important ingredient in training a senior programmer is the selection of the bright, apt, and interested student with an adequate academic background. Capability was far more important than was prior training (Howat, 1964). The computation center gains most from selecting and culling out such students as early as possible to benefit as long as possible from their experience. Using student programmers at a number of levels appears the best compromise to attain the computer center goal of efficiency and the educational goal of widespread training. Supervision which guides progress in the computing field, and interaction with colleagues in a warm and nourishing atmosphere of cooperation to maintain interest, rapport, and adequate communication should follow as indicated below.

Except as a consequence of D 's restricted time available for the Computer Center, all trends should have been continuous across time. Rapport with D was (and should be) sufficiently close to enable weekly expression of difficulties that S s could objectively discuss. The consequence of limited supervisory time is interpreted as the causal factor in W 's being rated as the quarter in which (1) the least was accomplished, (2) communication with D was the poorest, and (3) manner of job assignment least corresponded with S s' preference. Little useful work was accomplished in the F quarter compared to the W quarter, but S s rightfully include learning. The D sensed that more time for supervision during the W quarter would have been profitable. There is no substitute for sufficient supervisory time, readily available and properly used.

Along with administrative aptitude, the third resource needed is a sufficient supply of problems, challenging and yet suitable to the S s' specialization and skill. Although the activity of computer operating was an enjoyable task at the beginning of the year, by the end most S s preferred programming and consulting on the analysis. Since the problems that arise cannot always match interests, there is an additional value in having available some experienced and versatile programmers as well as operators on the permanent staff. It is safe to conclude that a capable programmer's greatest pleasure derives from solving computational problems in which he finds an intrinsic challenge. Such satisfaction implies allocating problems matching the programmer's proficiency.

A number of conditions must exist to ensure the maintainance of interest. A cooperative atmosphere is considered the most important basis for encouraging interests and rapport of the *Ss* working on closely-related task objectives. It gives essential recognition for individual differences. A convenient library of computing literature, small problems to practice new programming skills, time for work on problems of their own choice, encouraging the entry of a useful program in a cooperative library, and slack time for systems problems, may all contribute to maintaining interest. Each *D* will decide how many factors to include and how much to weight each.

Finally, two conditions demand that "quality of research throughput" (QUORT) be stressed. These are (1) easy access to computers, and (2) the frequency with which programmers have an opportunity to suggest alternative analyses when consulting directly with the users. Thus, for each original problem the proposed analysis should be reviewed by independent statistical or mathematical consultants. Such consulting or advisory services should occupy a parallel position with each research computer center.

REFERENCE

- Howat, M. G. "The Graduate Assistant in a Computer Training Position." Paper presented at Fall Joint Computer Conference, San Francisco, 1964.

BOOK REVIEWS

WILLIAM B. MICHAEL, Editor

University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor

Centinela Valley Union High School District

and the

University of Southern California

<i>Krathwohl, Bloom, and Masia's Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook II: Affective Domain.</i> JERRY S. WIGGINS	895
<i>Frederiksen and Gulliksen's Contributions to Mathematical Psychology.</i> DAVID M. MESSICK	897
<i>Green's Digital Computers in Research: An Introduction for Behavioral and Social Scientists.</i> M. GORDON HOWAT	900
<i>Davis' Educational Measurements and Their Interpretation.</i> ELLIS B. PAGE AND HERBERT GARBER	902
<i>Horrocks' Assessment of Behavior.</i> LEWIS R. AIKEN, JR.	904
<i>Gronlund's Measurement and Evaluation in Teaching.</i> QUENTIN C. STODOLA	907
<i>Engelhart's Improving Classroom Testing.</i> NORMAN C. MABERLY	909
<i>Engelhart's Improving Classroom Testing.</i> JULIAN C. STANLEY	911
<i>Ahmann's Testing Student Achievements and Aptitudes.</i> HENRY F. DIZNEY	911
<i>Sax's The Construction and Analysis of Educational and Psychological Tests: A Laboratory Manual.</i> RICHARD E. SCHUTZ	914
<i>Garrett's Testing for Teachers (Second Edition).</i> CARMEN J. FINLEY	915
<i>Yamamoto's Experimental Scoring Manuals for Minnesota Tests of Creative Thinking and Writing.</i> R. C. PAXSON AND R. W. BOYCE	917
<i>Hawes' Educational Testing for the Millions.</i> HENRY KACZKOWSKI	920

<i>Young and Veldman's Introductory Statistics for the Behavioral Sciences.</i> PETER A. TAYLOR	922
<i>Baggaley's Intermediate Correlational Methods.</i> JAMES A. WALSH	925
<i>Amos, Brown, and Mink's Statistical Concepts: A Basic Program.</i> STEPHEN W. BROWN	927
<i>Astin's Who Goes Where to College?</i> JULIAN C. STANLEY	927
<i>Travers' An Introduction to Educational Research.</i> GRETCHEN BRIEGLER TIMMERMANS	931
<i>Ong's The Opposite-Form Procedure in Inventory Construction and Research.</i> JULIAN C. STANLEY	933
<i>Ford and Pugno's The Structure of Knowledge and the Curriculum.</i> O. L. DAVIS, JR. AND JOHN M. KEAN	933
<i>Festinger, Schachter, and Back's Social Pressures in Informal Groups.</i> EDWARD LEVONIAN	938
<i>Bugelski's The Psychology of Learning Applied to Teaching.</i> HENRY KACZKOWSKI	940
<i>Boring's History, Psychology, and Science: Selected Papers</i> (Edited by R. I. Watson and D. T. Campbell). HAROLD BORKO	942
<i>Gowan and Demos' The Education and Guidance of the Ablest.</i> WILLIAM COLEMAN	943
<i>Passages from the "Idea Books" of Clark L. Hull</i> (Compiled by R. B. Ammons and Ruth Hays.) WALTER C. STANLEY ..	945

Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook II: Affective Domain by David R. Krathwohl, Benjamin S. Bloom, and Bertram B. Masia. New York: David McKay Company, 1964. Pp. vii + 196.

Since an APA meeting held in 1948, an eminent group of psychologists and educators have devoted considerable thought to the development of a conceptual framework through which the goals of education could be viewed with clarity. The immediate impetus for this effort was the recognition that a lack of common terminology and framework seemed, in large part, responsible for the difficulty encountered in translating educational goals into end products which could be objectively evaluated. It was felt that a standard taxonomy of educational objectives would facilitate communication, not only among those concerned with evaluation, but also among all professionally concerned with the educational process.

Working within the admittedly traditional trichotomy of cognitive, affective, and psychomotor behaviors the efforts of collaboration first bore fruit in the form of *Handbook I: Cognitive Domain*. Granting the less than finished nature of this product, it would seem unfair to view it as less than a milestone in the history of educational evaluation. The Cognitive Domain Taxonomy orders the realms of knowledge, abilities, and skills by reference to the single underlying principle of complexity. Such an ordering which appeared to be both sensible and useful, has been received with both praise and constructive suggestions regarding details. Spurred on by the reception of the cognitive taxonomy, the working group turned its attention to the formidable affective domain, the subject of this review.

As one might anticipate, a domain which includes interests, attitudes, values, appreciation, and adjustment cannot be dimensionalized with the degree of consensus possible for the better understood and more heavily emphasized domain of cognitive behaviors. Systematic evaluation of the attainment of affective objectives has heretofore been noticeably absent in practice. This absence may be attributed, in part, to the reluctance of educators to probe into matters which our culture considers "private" and to a concern with the distinction between indoctrination and education. There is also a widely held conviction that such objectives are attained only after

a considerable period of time. The authors, who would like to reopen some of these issues, feel that they are best discussed in terms provided by a sensible affective taxonomy. Another reason for hesitation in using affective measures for evaluation stems from "... the ease with which a student may exploit his ability to detect the responses which will be rewarded and the responses which will be penalized." This issue is rather blandly dismissed by the authors as a "technical problem" in the absence of any concrete suggestions for its solution.

It is proposed that the complex and heterogeneous domain of affective behaviors be organized with respect to an underlying central process called "internalization." By this is meant the developmental sequence whereby a phenomenon, characteristic, or value (which is initially not part of a student's affective repertory) becomes a meaningful organizing principle for the student. When the internalization process is viewed as a hierarchical continuum, several points on the continuum are recognizable as familiar psychological processes which have been studied in somewhat different contexts. At the lowest stage or level of the continuum the relationship between the student and the value is simply one of awareness or at best controlled attention to the value. At the second stage a development is noted from mere acquiescence to a willingness to respond and a satisfaction in responding to the affective stimulus. During the third stage the value is accepted, a preference for the value emerges, and a sense of commitment to the value is present. At the fourth stage the value is conceptualized as a value and accorded a place in the organization of other dominant values which determine the course of the student's life. At the fifth and highest level of internalization, the student develops a generalized set or world view, based on the value, which eventually may become so encompassing that the value is what uniquely characterizes the student.

Classification is facilitated by a detailed manual which provides three kinds of definition for each of the categories. The first definition is in the form of a concise verbal statement of the behavior implied by the category. The second definition takes the form of a listing of educational objectives culled from actual courses or curriculum program statements. The third definition consists of illustrative measures modified from previously developed test items.

An attempt is made to relate the taxonomy to more commonly used affective terms as well as to the developmental stages of such writers as Peck and Havighurst and Erikson. A detailed examination of the relationship of the affective to the cognitive domain reveals a close correspondence at the upper and lower levels. The relations between the two domains are seen to have implications for curriculum, evaluation, and research. From a formal point of view,

such correspondences between the two domains may bring into question the value of an initial dichotomy between cognitive and affective behaviors.

Viewed as a substantive contribution to a general theory of affective behavior, the present taxonomy falls short of a rigorous integration of the many concepts which it subsumes. Viewed as a starting point for more realistic consideration of the place of affective behaviors in curriculum, evaluation, and research the taxonomy is a notable achievement. This work seems likely to provoke considerable discussion and reexamination of previously obscure educational issues. This was, of course, its purpose.

JERRY S. WIGGINS
University of Illinois

Contributions to Mathematical Psychology by Norman Frederiksen and Harold Gulliksen (Editors). New York: Holt, Rinehart, and Winston, 1964. Pp. 189., \$6.50.

This book consists of seven papers which were read at a conference on mathematical psychology held by the Educational Testing Service in the spring of 1962. The occasion for the conference was the dedication of a new building, Thurstone Hall, named in honor of the late Louis Leon Thurstone. The contributors to the conference were selected both on the basis of the outstanding nature of their research and also because of the noticeable influence, direct or otherwise, of Thurstone's ideas in the development of their work.

The diversity of interests displayed by the papers in this volume attests to the breadth of Thurstone's influence in modern psychology. The quality of the work is a reflection of the soundness and originality of his ideas. Professor Thurstone will certainly be judged one of the truly great and original psychologists of the first half of this century, and this distinguished collection of papers provides an appropriate tribute to him.

The individual papers will be reviewed in order.

"Louis Leon Thurstone: Creative Thinker, Dedicated Teacher, Eminent Psychologist" by Dorothy C. Adkins.

This biography (which is followed by a complete bibliography of Thurstone's published and unpublished works totalling 372 entries) provides an interesting and insightful sketch of Thurstone's life and interests. We are shown how the seeds of his interests in psychological issues which were to bear fruit later were planted when Thurstone was an engineering student at Cornell. It is of interest that his penchant for mathematics was evident even before going to college. His career, which began as an assistant to Thomas A. Edison, is fascinating and often instructive. The picture which

emerges from this paper is that Thurstone was indeed a brilliant and dedicated person who was all of the things mentioned in the title.

"Some Symmetries and Dualities among Measurement Data Matrices" by Clyde H. Coombs.

The reader is taken on a relaxed tour through Coombs' Theory of Data. One will pause frequently to take note of the relationships existing between various types of scaling procedures which appear in different octants of the classificatory scheme. This is done with respect to three distinct types of data matrices; dominance matrices (in which the ij^{th} entry reflects a dominance relation), symmetric proximity matrices (in which the ij^{th} entry is a measure of the similarity of the i^{th} and j^{th} stimuli), and conditional proximity matrices (in which the ij^{th} entry is an estimate of the conditional probability of response j given stimulus i). It is shown that for each type of matrix, if one takes an off-diagonal submatrix from the square matrix, the model for the data in the submatrix will be different from that appropriate for the square matrix. Coombs concludes with a complete analysis of an example of a conditional proximity matrix.

"Intercultural Studies of Attitudes" by Harold Gulliksen.

Gulliksen presents some results stemming from a large intercultural study of attitudes toward occupations, nationalities, reasons for working, and goals of life. The data were obtained from students in Belgium (both French and Flemish), Italy, Norway, Germany, France, Texas, and Pennsylvania. Average scale values were computed for each item in each of the four schedules, and scatter plots of these scale values for various pairs of national groups are given. There is remarkable agreement among the groups with regard to the variables included in the study. The correlations which are presented range from .627 to .988.

The data matrix, incorporating all of the items in the four schedules, was then factor analyzed to yield the dimensions of preference for each group. Only data from the Belgium and Texas samples are presented. Again the most remarkable finding is the similarity in factor structure (even in cases where the scale values may differ considerably) between the groups. It will be of some interest to see whether the similarity of factor structure holds up when more widely discrepant cultural groups than those included in the present study are compared.

"The Extension of Factor Analysis to Three-Dimensional Matrices" by Ledyard R Tucker.

In this rather weighty paper a model for the extension of factor

analytic techniques to three dimensional matrices (e.g., judges \times attributes \times objects) is explored. The model is composed of three two-dimensional matrices relating the underlying variables to the data, and a three-dimensional core matrix which specifies the relations among the underlying variables. The effects of transformations within the model are investigated and a procedure analogous to principal components analysis is described for the solution of a data matrix for the structural matrices in the model.

"Matrix Factoring and Test Theory," by Paul Horst.

This article is addressed to the general problem of evaluating the dimensionality of a set of test items of varying degrees of difficulty. After mentioning previous work which has been done on the problem, Horst decides it might be best to attempt to incorporate the unequal dispersion phenomenon in a model to provide an explicit means for systematically accounting for it. This is done by assuming a latent simplex for the set of items and then computing the regression of the data matrix on the matrix representing the latent simplex. The residual covariance matrix is then factored to yield the dimensionality of the set after the preference structure has been removed.

"Mathematical Models of the Distribution of Attitudes under Controversy" by Robert P. Abelson.

The problem with which this paper deals is that of finding a reasonable dynamic model for the attitudes of group members during a controversy. The first model considered describes the rate of change of one's attitude on some issue as being proportional to the weighted sum of the discrepancies between that person's attitude and those of the others in the group, where the weights are the rates of contact with the other members. It is shown that if there is at least one person in the group who can influence, directly or through contact with others, everyone else in the group, then the only stable distribution of attitudes within the group is complete agreement. After a discussion of factors affecting the rate of convergence on the equilibrium distribution and the analysis of an example, an effort is made to extend the model. However, even for very general assumptions it is found that for "compact" groups, as defined above, the only stable distribution of attitude positions is complete agreement. Some alterations of the assumptions are discussed which could lead to bimodal or multimodal equilibrium distributions. Abelson concludes by giving a condensed description of a computer program constructed to simulate attitude modification during a group controversy.

Some New Looks at the Nature of Creative Processes" by J. P. Guilford.

The processes involved in creative thinking are considered in this paper from the point of view of the author's theory of intelligence, the "structure of intellect" model, which is a sort of Periodic Table of mental abilities. Abilities having to do with divergent production operations appear to play a major role in creative behavior. Abilities such as ideational fluency, associational fluency, spontaneous flexibility, and adaptive flexibility, the ability to vary one's approach to a problem, all appear to contribute at this level to creative and original thought. Another important factor seems to be the ability to detect the existence of problems, to recognize incongruities and imperfections.

While the approach to creativity presented here tends to be somewhat taxonomic, the author shows that it is related to other more dynamic conceptions of the creative process.

DAVID M. MESSICK
University of California
Santa Barbara

Digital Computers in Research: An Introduction for Behavioral and Social Scientists by Bert F. Green. New York: McGraw-Hill, 1963. Pp. viii + 333. \$10.75.

This book is the second to cover computer applications to behavioral sciences. The first, edited by Borko (1962), is a multi-authored book providing more detailed and advanced coverage of a number of topics. However, Green has achieved the advantage to be expected of a single-author book: a book at a uniform level designed as an introduction to behavioral application in a very readable and stimulating style.

In Part I a hypothetical programming language is used which assumes a hypothetical, although typical, computer. This approach is a curious one. It is obvious to any psychologist that use of a real computer and a live language will facilitate learning because of higher incentive value and quickly available knowledge of results. Any student who can profit from knowledge of how to use the computer as a tool certainly should take a regular programming course, or be given the needed guidance in learning to use a real computer.

Part II has two chapters: "Digital Codes" and "Programming for Information Processing." These topics also can be best treated in a regular programming course.

Part III, "Behavioral-Science Applications," pages 140 to 274, is the only portion of the book that justifies the reading attention of certain psychology or human factors students. Of the six chapters included not more than two are judged to be of interest to other social scientists. If the topics of Part III are inviting, then their

reading should create further interest in the application of computers to stimulus generation and control, to man-computer systems, to modeling or simulation, and to retrieving source information via the computer. One notable omission is a discussion of teaching machines and programs. It is regrettable that Green often misses opportunities to go deeper once a problem or need for an explanation is exposed. For example, the concluding statement of Chapter II, which refers to the simulation of group decisions, reads, "It seems that a computer program is the best, if not the only way, to formulate a model of this kind of behavior." Green does not offer to support this statement and thus leaves the reader to guess the reason.

Part III is not organized around principles of behavior or research methodology, but primarily around program capabilities. One does not find in one place, for example, all the considerations involved in the choice of simulation over live experimentation. Since it is an introduction and an overview, it does not explore the possibilities or principles involved in application at a sufficiently penetrating depth to justify its study by specialists. Such individuals should read the original literature.

Green merely introduces the use of computers for statistical analysis. This approach is sufficient because the topic is comprehensively treated elsewhere, e.g., in Cooley and Lohnes (1962). Because statistical analysis concerns all who will work with variable data, this application of computers should be included in the curriculum before any class time is devoted to the other topics of Part III, which few students will actively use. If the principles underlying such topics are used, most likely they will be employed as a research speciality. This book will be properly appreciated only after the programming capability is understood. At this point, the book will be easy reading. These arguments comprise the reasons why Parts III and IV are best used as reading references.

Part IV contains two chapters for the curious: Chapter 14 on how circuitry performs the computer operations, and Chapter 15 on Turing machines and a computability theory. Whether or not these topics should be included in a regular programming course or assigned to reference reading is up to the instructor's judgment.

In spite of suggestions to the contrary, including those made on the jacket blurb, no support can be found for the notion that this book should be used as a text, aside from the fact that provocative problems are found at the end of the chapters. This reviewer feels that the author should have restricted the book to the material in Part III, which is 44 percent of the main text, and should have published it at an appropriate fraction of the price asked, \$10.75. Nevertheless, the book should be readily available as a reference for readers interested in computer applications or experimental psychology.

REFERENCES

- Borko, H. (Ed.) *Computer Applications in the Behavioral Sciences*. Englewood Cliffs, N. J.: Prentice-Hall, 1962.
- Cooley, W. W. and Lohnes, P. R. *Multivariate Procedures for the Behavioral Sciences*. New York: Wiley, 1962.

M. GORDON HOWAT

*Cornell Aeronautical Laboratory of
Cornell University*

Educational Measurements and Their Interpretation by Frederick B. Davis. Belmont, California.: Wadsworth Publishing Company, 1964. Pp. ix + 422.

When an outstanding professional leader writes an introductory textbook, it is always an important event. Fortunately, such events have happened before in educational measurement; authors such as Julian Stanley (*Measurement in Today's Schools*, 4th ed. Englewood Cliffs, N. J.: Prentice-Hall, 1964) and Thorndike and Hagen (*Measurement and Evaluation in Psychology and Education* 3rd ed. New York: Wiley, 1961), and others have supplied authoritative and widely-adopted textbooks for the field. Now another text has been added to around a dozen available measurement books, and the newest would be interesting, if for no other reason, because Frederick Davis wrote it.

But there are other good reasons to be interested in Davis' new book which, according to the Preface, is "intended as a basic text for courses in psychological and educational measurement and as a handbook for clinical and school psychologists, counselors, and school administrators" (p. vii). If this undertaking sounds too diffuse, Professor Davis has avoided failure and indeed has managed a formidable task with judgment and skill.

The book is modular in construction, with certain pieces which may be almost independently employed. To give the practical flavor of the book, the 13 chapters are worth listing: (1) Measurement and its uses in the schools, (2) Basic characteristics of test scores, (3) Selection and administration of standardized tests, (4) Test scoring, (5) Measurement of achievement, (6) Measurement of intelligence and aptitude, (7) Measurement of interest, (8) Interpretation of individual test scores, (9) Interpretation of group scores, (10) Interpretation of change, (11) Interpretation of under- and over-achievement, (12) Constructing tests for classroom use, and (13) School marking procedures. Thirteen appendices (85 pages) independently treat such statistical matters as computing arithmetic means, percentiles, reliability coefficients, standard errors, and various tables.

This modular structure is intended by Davis, according to the

Preface. He regards one arrangement (chapters 1-8, 12, 13, and certain appendices) as comprising a basic course in measurement for upper division or graduate students, and regards a different arrangement (chapters 8-13 and certain appendices) as more helpful to advanced students and professionals in the field. Taken literally, such a division would result in a beginning text of around 275 pages plus appendices, and an advanced text of only around 137 pages, plus appendices and "review material." The suggestion of much elementary material as necessary review for many advanced students is, alas, probably all too realistic.

One of the author's main purposes was to "supplement rather than duplicate" the traditional textbooks in the field. Here is meant the large amount of attention to the "interpretation of individual and group scores, the measurement of change, the measurement of over- and underachievement, and school marking procedures" (p. viii). These sections are indeed unusual and valuable for the practical educator, and contribute much to the book's usefulness. Among commendable qualities is a consistent functionalist orientation, which perhaps comes through most clearly in Chapter 6, where the author demonstrates that to define aptitude one necessarily relates a test score to a specified criterion. Another is found in Chapters 8 and 9, in rational admonitions to avoid unjustified interpretations of data. And in Chapter 10, in the treatment of the measurement of change, regression is made decently understandable to the beginner yet not too disturbing to the initiated. And there are many other virtues.

Certain shortcomings are perhaps the obverse side of the virtues. The intention is so ambitious and the format so narrow (422 pages) that one thinks of band width and fidelity. The book is dense with useful information, to the point of being perhaps inadequately redundant, and therefore less than ideal pedagogy. And at present, there is regrettably little help to the teacher in review, discussion, or testing material. A promised companion volume, a self-teaching workbook not available at review time, may remedy any fault here, if fault there is, and lend the book more pedagogical convenience. Of course, for certain of Professor Davis' purposes, beginning students are not the target audience, but these limitations should be noted for the instructor whose students need application of knowledge and a more complete instructional loop.

Also, certain emphases in this first edition may be objected to by instructors of more advanced courses. Some portions seem, rather puzzlingly, not in step with the times. For example, the fourth chapter makes much of an alleged superiority of correcting for guessing compared with number-right scoring. Is such an emphasis justified? Many measurement professionals may believe the issue is not worth the attention given, and many also may disagree with

the conclusion. Many would argue that, in general, students should be urged to "guess" whether there is a typical scoring correction or not, and research presented by Davis himself (pp. 84-85) indicates that it usually pays. The reviewers know of no hard evidence that students who guess without penalty become "cynical and contemptuous" (p. 90), and other recent writings flatly contradict Davis (cf. Robert L. Ebel, *Measuring Educational Achievement*, Englewood Cliffs, N. J.: Prentice-Hall, 1965, pp. 223-233.) The effort would be better spent teaching school people to avoid speededness in most testing situations, with the result that the whole question of guessing becomes less important.

Another anachronism may be the failure to discuss advanced technology. There is reference to the hand-fed IBM 805 which, of course, even at the school level, is rapidly becoming a museum piece. In a book stressing the *interpretation* of test scores, surely there should be some treatment of computer scaling and norming, and of combination of scores and other measures for predictive or classification purposes. Here there is not. There is little or no treatment (and no indexing) of computers, multiple regression, multivariate analysis, profile analysis, or most other improved uses of scores made possible by expanding computerization. Professor Davis (who was recently Director of Project Talent and who of course knows much of such advances) must have decided that such discussions were not yet appropriate to the school counselor, and that the point of view should remain essentially what it was 20 years ago. Surely, most competitive books are also weak (in the opinions of the reviewers) in this way. But it may be hoped that the next round of text revisions will take cognizance of the technological counseling and instructional assistance which is becoming increasingly available. Measurement specialists should not shackle the next generation of school people to an outworn and limited technology.

But these shortcomings (compared with an elusive ideal) should not dominate this review. Professor Davis has inevitably contributed to measurement an important book, which will be read with interest by many and adopted by some. Whether it serves better than other works for any particular need must be decided after careful individual study.

ELLIS B. PAGE AND
HERBERT GARBER
University of Connecticut

Assessment of Behavior by John E. Horrocks. Columbus, Ohio: Charles E. Merrill, 1964. Pp. xv + 736.

The writer of a textbook on educational and psychological tests is always faced with the problem of how much statistical theory to

include in his book. On the one hand, he may choose to emphasize the statistical bases of test theory and thus perhaps give very brief descriptions of specific tests as illustrative applications of the theory. On the other hand, he may feel, as Professor Horrocks does, that space in a measurement text is too dear to include a course in elementary statistics.

This book, which bears the subtitle "The Methodology and Content of Psychological Measurement," is certainly no statistics book. "Mathemaphobes" will be comfortable with the book, and tests and measurements instructors with an abundance of these students will do well to consider adopting it. The book is best characterized as a "comprehensive survey" of available psychological tests, including discussions of the history and nature of psychological measurement. Obviously a product of painstaking scholarship, it represents the collaborative efforts of Professor John Horrocks and the late Dr. Winifred B. Horrocks.

The book is rather long. Its length may thus be a disadvantage at a time when the trend in textbooks appears to be toward the short paperback or compendium. However, the book is written in a very readable style, and there are exceptionally few typographical errors for a first edition. Numerous drawings, photographs, and charts are used to illustrate and to describe principles and tests.

Some of the reviewer's reactions to the 20 chapters comprising the book are as follows. In the first chapter, "Measurement in the Science of Psychology," the use of the phrase "method of measurement" rather than the more customary "method of correlation" is somewhat confusing. Although the author apparently means more than correlation by "method of measurement," exactly what is meant is not clear. In addition, it is the opinion of the reviewer that a more appropriate conception of the role of measurement in science is that of an ancillary technique rather than a "method."

Chapter 2, "Individual Differences and Measurement," is a summary of the characteristics of distributions and standards of measurement. But the author's coverage of these statistical topics is accomplished with words and illustrations rather than with formulas. Characteristic of this approach is the fact that the statistical formulas given are placed in footnotes.

Chapter 3, "Attributes of a Measuring Instrument," is a brief discussion of validity and reliability and of certain external criteria for judging tests. Again, the author uses words rather than algebra. In contrast to the usual way of approaching this subject, there is very little discussion of correlation; standard errors of estimate and measurement are not even mentioned.

Chapter 4, "Measuring Instruments: Classification and Sources," contains useful information on methods of classifying tests, sources of test information, and the ethics of testing. The descriptions of

systems of coding tests and of organizing a test library are of particular interest, since they are not commonly included in books on psychological measurement.

Chapter 5, "The Nature and Meaning of Intelligence," is very comprehensive in its discussion of mechanistic and abstractive approaches to intelligence. The three succeeding chapters survey individual intelligence tests, including Binet-type tests (Chapter 6), performance scales (Chapter 7), and Wechsler scales (Chapter 8). Although the author is rather mild in his criticisms of some of these instruments, he does emphasize the inadequacy of standardization of the majority of them.

Chapter 9 gives numerous examples and illustrations of group intelligence tests, and Chapter 10 is a good summary of the designs and purposes of several so-called "culture-free" tests.

Chapter 11, which is a general chapter on the rationale and history of aptitude testing, features a review of Guilford's scheme for classifying aptitudes. Chapter 12 is a summary of available aptitude tests.

Chapter 13, which discusses infant intelligence tests such as those by Gesell and by Cattell and Griffiths, is followed by Chapter 14 on test of subject-matter readiness. Incidentally, the majority of the typographical errors in this book detected by the reviewer are in Chapter 14. Chapter 15 is on the measurement of achievement.

Chapters 16 to 19 discuss personality measurement. Chapter 16, "The Nature and Meaning of Personality," does for the concept of personality what Chapter 5 does for intelligence, through summarizing approaches to defining and measuring personality. Chapter 17 surveys personality inventories, and Chapter 18 is a fairly "optimistic" account of personality rating instruments and techniques. The discussion of projectives in Chapter 19 illustrates the quality of scholarship which is evident throughout the book. This chapter is a thorough and critically restrained discussion of the methodology and content of projective measures of personality.

Chapter 20 is a summary of measures of interests, attitudes, and social behavior.

Some of the unique features of this volume may be considered as assets and liabilities. On the positive side of the ledger are the careful, scholarly reviews of the most extensively used standardized tests and the complete lists of tests in each area. In addition, the chapters on the nature of intelligence and personality serve to orient the reader in his attempt to understand both why particular tests were constructed and which types of tests might still be devised. Finally, the minimal statistical content of the volume will endear the author to those with little interest or preparation in statistics—an audience which may well encompass the majority of students enrolling in tests and measurements courses.

On the negative side, the book has many of the characteristics of a reference work rather than a text. It contains long, detailed discussions of specific tests, and this feature may tend to make the book uninteresting at times to certain members of a student audience. Finally, the omission of much important statistical material leaves the instructor who adopts this text with a problem. Although a course in psychological statistics may facilitate comprehension of the text material, it is not essential. Therefore, should the instructor supplement this text with material from a more quantitatively oriented book on psychological measurement, or should he be satisfied with discussing only the statistical material in the Horrocks text? If he chooses the latter alternative, his course will probably be more popular, but will he be teaching all of the important aspects of the methodology and content of psychological measurement?

LEWIS R. AIKEN, JR.
*University of North Carolina
at Greensboro*

Measurement and Evaluation in Teaching by Norman E. Gronlund. New York: Macmillan Co., 1965. Pp. vii. + 420.

Recently several textbooks have been published on classroom evaluation procedures. Now that Dr. Gronlund has added his entry to the list, it seems only fair to ask: Is another measurement textbook of this type really needed? The answer to this question, in the reviewer's opinion, is definitely "yes." Dr. Gronlund's book, *Measurement and Evaluation in Teaching*, presents an exceptionally comprehensive and well-integrated statement summarizing much of the best thinking of the past few years on pupil evaluation. Straightforward in presentation, the book should be clearly understood by elementary and secondary school teachers, as well as by school administrators, supervisors, and counselors.

Dr. Gronlund emphasizes that the major purpose of evaluation should be to improve learning. Consequently, much of his discussion on the use of classroom tests, standardized tests, and observation techniques is devoted to explanation of how evaluation may assist in the learning process. As he clearly states, the most important reason for testing is to obtain a measure of the present level of pupil achievement in order to provide a meaningful basis for planning instruction.

Particularly noteworthy is the section on construction of classroom tests. This section, which is more than 100 pages in length, is very well organized. A carefully developed description of different item types is offered with clear-cut explanations of how each may be used for measuring a variety of behaviors. These item

types range from short answer items to interpretive exercises. The sample items used for illustrative purposes are excellent.

There are other areas of strength in the book. There is a well organized discussion of the need to define teaching objectives and to relate these carefully to evaluation procedures. It should be mentioned also that the suggestions for further readings are especially good. They are carefully selected so that they should prove both interesting and useful for teachers.

If this book has a weakness, it is one that is shared in common with most other recent books on pupil evaluation. It accepts without critical comment much of the "folklore" of modern measurement, even when the principles and procedures advocated may be impractical or even undesirable for classroom teachers to follow.

For example, Dr. Gronlund advocates the use of items which are at the 50 percent level of difficulty in order to obtain reliability in measuring achievement—i.e., he wants to obtain the largest possible spread among pupil's scores. In this connection he states, "In measuring the extent to which pupils are achieving our course objectives, we have no absolute standard by which to determine their progress. A pupil's achievement can be regarded as high or low only by comparing it with the achievement of other pupils" (p. 112).

Despite Dr. Gronlund's statement, and those of some other measurement specialists, many classroom teachers will continue to follow grading practices based upon comparing a student's achievement with standards set by the teacher himself rather than with standards reflecting relative achievement within a group. In this case, contrary to measurement dogma, the reviewer believes that teachers can and generally should set arbitrary standards of classroom achievement not based on a normative approach. Item difficulty is then determined by how well pupils meet teacher standards, not necessarily by how well the level of item difficulty increases reliability.

Dr. Gronlund further advocates that teachers should read reviews in Buros' *Mental Measurement Yearbooks* and also examine reliability and validity data provided in test manuals—all this as an essential part of the process of evaluating standardized achievement tests. From the reviewer's experience teachers are simply not able to absorb the necessary technical background in a single course in tests and measurements so that they can fully comprehend many of the test reviews in Buros' *Yearbooks*. Nor can they make much sense of the kinds of validity and reliability data ordinarily provided in test manuals. The reviewer believes that the best way for classroom teachers to evaluate standardized achievement tests is almost entirely through content analysis.

It is recognized, of course, that many measurement experts will

regard Dr. Gronlund's support of traditional principles and procedures in a favorable light and that they will not share the reviewer's sense of uneasiness about approval of certain generally accepted ideas in tests and measurements. All in all, it should be stated that this is a good textbook. No doubt, many instructors in tests and measurements will choose to use it in their classes.

QUENTIN C. STODOLA

Director of Testing

Northern Michigan University

Improving Classroom Testing by Max D. Engelhart. What Research Says to the Teacher, No. 31. Washington, D. C. 20036: National Education Association, 1201 Sixteenth Street, N.W., 1964. 33 pp. \$0.25 per single copy. Quantity orders at the following discounts: 2-9 copies, 10 percent, and 10 or more copies, 20 percent, postpaid if payment accompanies order. A joint publication of the NEA Department of Classroom Teachers and the American Educational Research Association.

Widespread increase in the use of standardized tests has offered, one would hope, a substantial opportunity for improvement in many evaluation procedures formerly carried on by courageous teachers with their limited resources and knowledge. Nevertheless, despite the relief afforded teachers by standardized programs, classroom testing is still the very heart and core of pupil measurement. Daily quizzes and exercises, midterm and final examinations, oral questioning, and sundry other evaluative activities are such a vital part of teachers' work that it is something of a mystery why so many books on measurement give so little emphasis to specific recommendations and suggestions which might enable teachers to improve their testing techniques. In many instances, teachers' knowledge of esoteric measurement concepts is directly related to the degree of confidence that may be placed in the results they obtain; yet, search as they will, the prospects of finding understandable and adequately practical assistance are very remote indeed.

Although Engelhart's little treasure is hardly a full answer to the classroom teacher's prayer, it certainly is a step in the right direction. A refreshing aspect of the booklet is that such a knowledgeable expert is able to convey meaningful essentials in a way that should be understandable for even the most unsophisticated novice in classroom testing.

The body of the booklet is largely devoted to brief yet adequate discussions on "Objectives and Test Making," "Essay Testing," "Semi-objective Items," and "Objective Testing," followed by a rather less adequate introduction to the analysis and use of test data, although within the limits imposed by a rigid specification of

32 printed pages the author has done as well as could be reasonably expected. It is doubtful that many teachers have either the time or the motivation to engage in the many desirable procedures recommended, but, as Engelhart points out, ". . . these things cannot be accomplished overnight." Nevertheless, teachers should be given as much encouragement as possible to make use of the procedures and recommendations presented. Perfection is an elusive thing to be attained through successive applications of goal-directed principles and techniques which may appear at first sight by teachers to be a little overwhelming.

Much confusion in learning is doubtlessly introduced through semantic limitations. A term such as "semiobjective" tends to connote a neither-nor conflict that must be resolved, ultimately, in one direction or the other. In common measurement parlance, it would seem that a test item is classified as "objective" if it may be reliably scored by any number of independent scorers. An item may be classified on the other end of the continuum if the scoring procedure is subjective in the sense that it does not yield a high degree of scorer reliability. The suggestion that completion and short-answer types of items are neither objective nor subjective, but somewhere halfway between, is a beclouding concept of dubious worth. The lack of objectivity in such items often arises through the inabilities of the teacher to phrase the items in non-ambiguous terms. What the teacher really needs is more advice on how to overcome the pitfalls that destroy the objectivity of such items.

Although there has been criticism of many so-called objective item-types, including Engelhart's examples, experience and research have demonstrated that knowledge and understanding, as well as many other important outcomes of learning, may be best evaluated through reliance on variations of the multiple-choice item. Perhaps this is why so much space in the booklet is devoted to many excellent suggestions and recommendations which should enable the discerning teacher to build a sound objective testing program. The justification for occasional use of unstructured tests and essay-type examinations is not overlooked, for they too play an important part in providing insight into seemingly less tangible outcomes of instruction.

Spirited rebuttals from teachers and test specialists alike are likely to be aroused by certain suggestions concerning the conversion of test scores to marks; but notwithstanding the inherent dangers of the percentage method proposed, there may be some consolation in the fact that this is one type of activity for which teachers are slow to take advice and for which they are quickly put on the defensive. Consequently, in the few places where they are likely to be put into practice, the suggestions should contribute to the development of marking systems that are considerably superior to

those often employed by many teachers whose marking methods have no supportable basis whatsoever.

Improving Classroom Testing is a booklet that needs to be placed in the hands of all teachers at all levels of education, including those at the college level, with the hope that it will stimulate a greater interest and understanding in the dynamics of testing, while at the same time it will provide a means of more equitably evaluating pupil progress than now exists. Test specialists and directors of in-service training should also find the booklet useful as a means of promoting and improving classroom testing in local school systems.

NORMAN C. MABERLY
Test Department
Harcourt, Brace & World, Inc.

Improving Classroom Testing by Max D. Engelhart. What Research Says to the Teacher, No. 31. Washington, D. C. 20036: National Education Association, 1201 Sixteenth Street, N.W., 1964. 33 pp. \$0.25 per single copy. Quantity orders at the following discounts: 2-9 copies, 10 percent, and 10 or more copies, 20 percent, postpaid if payment accompanies order. A joint publication of the NEA Department of Classroom Teachers and the American Educational Research Association.

In this attractive little booklet, an excellent way to begin the in-service training of classroom teachers and school administrators in measurement and evaluation, Dr. Engelhart draws on his long experience in the Chicago public schools to consider "Instructional Objectives and Test Making," "Essay Testing," "Semi-objective Items," "Objective Testing," "Converting Scores on Tests to Marks," "Item Analysis of Teacher-Made Tests," and "Using Test Data in Improving Instruction." Also, he provides 40 selected references.

Engelhart's treatment is more substantial than that which characterizes a number of the other booklets in this series. Educational practitioners can get a good overview of testing principles quickly and inexpensively here.

JULIAN C. STANLEY
Center for Advanced Study in the Behavioral
Sciences

Testing Student Achievements and Aptitudes by J. Stanley Ahmann. Washington, D. C.: The Center for Applied Research in Education, Inc., 1962. Pp. 118.

The foreword states that the book's purpose is to discuss applications of psychological tests for educational uses. As the title suggests, the emphasis is upon achievement and aptitude testing. Spe-

cifically, the book undertakes descriptions of (1) principles of test development, (2) methods of test construction, (3) test usage, and (4) the interpretation of test results. Generally, it attempts to assess the utility of tests in the educational process. As Eric Gardner says in the foreword, "This is indeed a major task."

There can be little doubt about the timeliness and appropriateness of such an undertaking. The field of education obviously represents the great testing market in terms of both absolute size and relative frequency of application and usage. Testing practices—good and bad—surely affect more people more frequently and in more fundamental ways in educational contexts than in any other. It is, therefore, unfortunate that a stronger case is not presented by this particular book. Its organizational structure and stated purpose promise much. Its development tends to be a compromise of information that does not fulfill that promise. To the reviewer, the basic weakness seems to be that the author has not clearly identified a reading audience for the book.

As technical material it lacks depth and accuracy. For example, concerning scales:

If well-developed psychological tests which tend to measure a single human characteristic are being applied, the data are generally thought of as falling within the interval level [of measurement]. (p. 5)

As a reason for non-ratio type scores:

In the interests of efficient testing, the authors of psychological tests design their instruments so that a zero measurement by the test does not conform to zero amount of the trait being measured. (p. 5)

In introducing the notion of relative ranking and norms:

On the basis of this casual inspection, [of a distribution of 103 algebra scores] it is anticipated that the distribution of raw scores may very well be approximately normal, that is, a bell-shaped or gaussian (*sic*) distribution. (p. 59)

On the problem of student guessing in objective-type tests, no appreciation of the limits for such scores in terms of rules of chance or for the positive appeal of effective foils is shown. Factor analysis seems to be reified when the distinction between its heuristic use and its use as a method for the verification of logically derived hypotheses is not made clear, e.g., "For a single test the test builder can discover [by factor analysis] the nature of the underlying psychological processes which determine the results yielded by the test." (p. 39)

As a textbook for instructional purposes, the reviewer's opinion is that it is too shallow in a field where outstanding texts are not unknown. The chapter on test reliability, validity, and norms is completed in 25 pages. This chapter is characterized by a great deal of talk about but rare operationalization of the topics at hand. Its two figures and two tables are inadequately tied to the discussion. Pearson's r is defined solely by its numerical limits and a scatter diagram. A paucity of formulas is given—none with adequate elaboration. There is a tendency throughout the book to justify procedures by popular practice instead of by explanation. In addition, there are passages so awkwardly written as to contribute to student confusion. For example:

It is surprisingly difficult to classify tests into types. In a number of instances fifteen important types have been identified, and these are not mutually exclusive types. A convenient way of examining the various types is to group them into two's or three's including in each group types which are more or less contrasting. The fifteen types have been classified into six groups.
(p. 6)

If the purpose of the book is to communicate to the general public, it seems apparent that it would only add further to the confusion existing in an already overly-charged area of concern.

On the positive side, Ahmann points out that the impact of testing on formal education, dramatic as it has been, could stand the expanded assistance of non-test evaluation instruments such as "ranking and rating devices, sociometric techniques, questionnaires, and anecdotal records." Unfortunately, he does not develop this theme. He does properly emphasize the primacy and inherent relevance of "paper-pencil" test situations to many educational goals.

To educators, the crux of testing must be evaluation. It seems to the reviewer that this book joins the majority of the literature in the quality of its discussion concerning the relationship between the act of assessment and the purpose of assessment. The problem is honestly introduced, but its treatment lacks specificity. In this case, the *Taxonomy of Educational Objectives* is "pasted" into the problem of the derivation of educational objectives and somewhat dogmatized, e.g., "The three classifications [cognitive, affective, psychomotor] represent the total framework of educational objectives." Meaningful procedures for making judgments with respect to the relative importance of the various categories of the taxonomy or to any other educational objectives are absent. No criterion other than the taxonomy (if it is a criterion) for the generation of educational objectives is presented. Educators may well wonder about the general lack of discussion in the measurement literature of

functional criteria (vague as they may be) such as social utility, ultimate truth, appropriateness to learner characteristics, or even efficiency.

It is unfortunate indeed that the stated purpose of Ahmann's book is not more nearly realized in itself. One could expect more on the basis of his earlier work in both measurement and statistics.

HENRY F. DIZNEY

Kent State University

The Construction and Analysis of Educational and Psychological Tests: A Laboratory Manual by Gilbert Sax. Madison, Wisconsin: College Printing and Typing Co., 1962. Pp. ii + 74. \$2.20.

Many measurement instructors would be favorably disposed toward the possibility of including formal laboratory experiences in their courses were adequate instructional materials available to support the laboratory activities. Sax has provided at least the nucleus of such materials. His manual includes exercises on 18 standard topics in an introductory course, ranging from "Reasons for Using Tests" to "Reporting Test Results."

Although the exercises are predominately paper-pencil based, they ingeniously tap aspects of performance that the reviewer regards as important components of the behavioral repertoire of students completing an introductory course. The exercises relating to test construction and item analysis are excellent. One is most impressed, however, with the materials relating to the interpretation and use of test results. Introductory measurement courses typically stress test usage as an objective, but do very little to equip the student with test usage skills. Unfortunately, many students seldom seem to achieve the transfer of training in this area which the reviewer in his teaching optimistically assumes that they should effect.

Sax presents exercises requiring interpretation of the test results of a fourth grade class and interpretation of two individual case studies. In these, the student is guided through a simulated experience of "how it's actually done." This experience should provide at least one empirical step toward accomplishment of the objective: "Uses test results to improve the adequacy of instructional decisions."

To facilitate use of the manual with the text of one's choice, Sax includes a table coordinating each exercise with the relevant page references in 14 currently available measurement texts. This table should also be useful in encouraging students to consult multiple sources in responding to the exercises.

Introductory measurement courses are often reputed to be "dry." Use of this manual will add a considerable amount of life to any in-

troductory course, with no loss of academic respectability and with a much higher probability that the students will accomplish the course objectives.

RICHARD E. SCHUTZ
Arizona State University

Testing for Teachers (Second Edition) by Henry E. Garrett. New York: American Book Co., 1965. Pp. vii + 280.

As a book written primarily for prospective teachers and with some eye to use as a guide for teachers in service, the author has made an obvious effort to simplify the sometimes heady content of texts in measurement and evaluation. The content is heavily weighted with classified descriptions of the various individual and group administered tests on the market today, with the usual preliminary historical introduction and essential statistical minimums. Case studies are presented in an attempt to illustrate the use and interpretation of test results. Three final chapters deal with teacher-made tests and with some case histories for evaluation by the student.

Except for the copyright date in Garrett's second edition of *Testing for Teachers*, this reviewer would have assumed a publishing date in the 30's or 40's.

The attempt which is made to relate test results to educational practices is not realistic in terms of present day trends. For example:

In Table 3-3, page 58, *Educational Expectation in Relation to IQ Level* the author states that for IQ ranges of 80-89 the child "... will complete the eighth grade—if at all—two or three years behind schedule." While at the 75-79 range, "These children may reach the fifth grade. Will rarely go beyond unless given much individual attention," and for below 75 IQ, "If one of these children reaches the fifth grade he will be fourteen-fifteen years old. Unable to do fifth grade work; but because of chronological age is likely to be pushed ahead after repeating each grade two or three times."

This reviewer submits that although these generalizations may have been valid a generation or two ago, they simply no longer exist in the majority of school practices today. An increasing number of youngsters at these lower ranges of ability are being educated in special classes, but many progress through regular grades with more than one retention being the exception rather than the rule. Although it is true that some will undoubtedly be "drop-outs," it is also true that some will graduate from high school.

The same unrealistic interpretation in the light of current curricular practices is noted in a number of the case studies presented.

For example:

Page 62, Case 3. *HP*, a boy; $CA = 6-5$, $MA = 9-6$, $IQ = 148$. "... *HP* is a very bright youngster. He should be ready for high school by age twelve or earlier. He should now be in the fourth grade, if he is ready for it socially." A mental age equal to that of the average third grader at one time was interpreted in this manner. However, research findings indicate the learning patterns of a $9\frac{1}{2}$ year old with a mental age of $9\frac{1}{2}$ is vastly different than a $6\frac{1}{2}$ year old with a mental age of $9\frac{1}{2}$ and in actual practice one rarely finds acceleration to the extent indicated.

Page 80, Case 1. Donald B.: age, 10-2; *WISC IQ*, 92; Arthur Scale *IQ*, 106. Under *Recommendation* Garrett states, "Donald's performance *IQ* is fourteen points higher than his *WISC IQ*. In view of his relatively meager abstract intelligence, this boy is probably doing as well as we can expect. He may get to high school, but will almost certainly not complete more than one year. Vocational training seems to be indicated. He will continue to have trouble with verbal subjects, but may be very successful at a skilled trade."

A number of gross over-generalizations are apparent here. First the reader does not know whether the *WISC* score is the *WISC Verbal* or *Total*, but assuming it is the *Verbal* (which would be the more extreme case) and taking into account the standard error of the respective tests, the evidence is marginal at best to indicate that these obtained scores really represent different true scores, i.e., that the discrepancy between them arose out of other than chance factors.

Again the prediction that he "will almost certainly not complete more than one year" is unrealistic according to present practices. The implication that if he cannot succeed in the academic world he will be successful in a skilled trade is not necessarily true. Such a point of view is the bane of the life of many vocational teachers. Further aptitude testing would seem indicated before any conclusion such as this one could even be suggested.

His discussion of the *IQ* as a mental age concept vs. the deviation *IQ* concept is confusing and contradictory. For example he goes into detail on the 1937 revision of the Stanford-Binet. Then in three short paragraphs he describes the 1960 edition and then states that this edition uses the deviation *IQ*. This approach is interesting when taken with his statement on page 55 that "The new deviation *IQ* represents a real improvement over the ratio *IQ*. However, the *IQ* as MA/CA has been associated with the Stanford-Binet Scale for nearly a half century and is firmly entrenched in the testing literature. Hence, the ratio *IQ* will probably continue to be used for a long

time. But it seems certain that eventually it will give way to the deviation *IQ*."

The trend away from the *MA/CA* ratio is readily seen in the newer editions of group tests; and with the change of the 1960 Binet to the deviation *IQ*, it would seem logical that few, if any, new individual tests of any import yet to be developed would return to the ratio *IQ*. On page 74 in contrasting the Binet and the *WISC* he states, "The *WISC IQ* is a deviation *IQ*—a standard score in a distribution with Mean = 100 and *SD* = 15—while the Stanford-Binet *IQ* is a development ratio or *MA/CA*."

Garrett, reminiscent of a 1937 debate with Guilford, still holds to his interpretation of the standard error of measurement as a measure of fluctuation around an obtained score rather than around a "true" score.

Probably the most helpful part of this book is in the listing and description of a fairly good selection of tests in use today. Such a cataloging is convenient to have in one place.

However, the interpretative material, in the opinion of this reviewer should be used with a great deal of caution and only by those who are aware of the discrepancies between the practices Garrett describes and what is actually happening in the schools today.

CARMEN J. FINLEY
Sonoma County Schools
Santa Rosa, California

Experimental Scoring Manuals for Minnesota Tests of Creative Thinking and Writing by Kaoru Yamamoto. Kent, Ohio: Bureau of Educational Research, Kent State University, May, 1964. Pp. 160.

As an advanced organizer, the reviewers point out that it was necessary to allude to the rationale underlying the various "Minnesota Tests" in sources outside the *Monograph* titled above. This is to say it was necessary to consult, for one example, Torrance (1962) in appropriately examining test administrative procedures and scoring schemes. The test user should be well informed about Torrance's theory for necessary background in appropriately interpreting the test manuals. The users of older, unpublished, mimeographed manuals of the test will be dealing with familiar materials with some modifications which are presented in a more condensed and organized form in the present manual than in the original sources.

In his "Acknowledgement," Yamamoto states that the sole purpose for publishing this *Monograph* is to make the "Minnesota Tests" available "for public examination and use" (p. 6). Furthermore, he requests feedback from potential users to improve upon

and to refine the various test manuals, but he does not explicitly state any limitations or restrictions in terms of his potential "public" or "users." Although this publication might be recognized as one of the successive approximations toward a more adequate administrative and scoring manual, possibly, there should be serious questions raised in terms of the content of the *Manuals* which will be seized upon as an absolute by many of those persons who need absolutes—precisely, those who are in a position to influence a child's education, but who lack the necessary background to do so. This action, of course, would be beyond Torrance's and Yamamoto's control—except to the extent that necessary qualifications for those using the *Manuals* should be made explicit.

Torrance states in the "Foreword" that "... a unifying rationale runs throughout the scoring of all tasks, verbal and non-verbal" (p. 8). Such attributes as *fluency*, *flexibility*, *originality*, and *elaboration* are included in the scoring scheme. There have been admirable improvements attempted by Yamamoto in elucidating the attributes mentioned above, one from the other. However, there are indications (admission in some cases) that the test designers have been preoccupied with *frequency of response* to the extent that other connotations of creativity may be leveled. It is possible that the weighting of responses such as those clustered under *originality* protocols, when accounted for in terms of frequency, could conceal spiral aspects of creativity. According to Yamamoto, this assessment, in terms of frequencies, would be the "... occurrence among an appropriate population" (p. 10). This orientation would mean that each user of the tests would need to determine the frequency of such responses in accordance with a local population in appropriately locating *high* and *low* level responses. This notion needs to be explicitly spelled out and repeated in key places in the *Monograph*, if sampling error is to be avoided. (Furthermore, what may be an uncommon response in one section or area of the country may not be so uncommon in another.)

It appears, upon close examination of the various tasks, that Torrance, et al., have devised a divergent-verbal-achievement test which may overlap only a small part of whatever is meant by their definition of creativity. In the reviewers' opinion, the subject, who is high in divergent-verbal achievement and who is achieving on a level where he is capable of organizing his thoughts as written statements is much better off on these tests than is a child who has not mastered the necessary skills needed in making adequate written statements. The *Monograph* fails to elaborate or to elucidate on possible difficulties faced by children. In this context, Torrance (1965, pp. 277-278) points out that a certain fourth grade boy demonstrated phenomenal success on the verbal tests when he was permitted to account orally for his ideas.

There are indications, although vague, that some children may provide evidence of higher levels of creativeness on the non-verbal tasks, while at the same time they score extremely low on the verbal written measures. The reviewers recommend that Yamamoto, et al., make these conditions more explicit. Further, the claims made by Torrance and others concerning the "slump in creativeness at the Fourth [grade] level" may need to be revisited, since there are reasons to believe that the lack of suitable achievement levels in basic educational skills may be inhibiting the writing of responses in an elaborate manner. Perhaps, the designers of the "Minnesota Tests" need an "error filter" in their system design.

The "Minnesota Tests" have had at least four revisions of the scoring scheme. These revisions resulted in more confusion than clarity. The present manual is an attempt to simplify a more systematic approach and thereby to achieve a more reliable and hence more objective system. The basic assumption that creativity (as outlined by Guilford) consists of *abilities* rather than one *ability* was accepted by Yamamoto. These abilities are *redefinition, sensitivity to problems, fluency and flexibility of ideas, originality, and elaboration*. From this theoretical position, Yamamoto extracted *fluency, flexibility, originality, and elaboration* as test protocols which would tap different aspects of creative thinking. This action was taken on an a priori basis ("... protocols which, at face value, would represent . . .") for "most tasks," but not all. The intercorrelations of the subscores indicate that this goal of differentiation of abilities was not achieved. On fifth-graders, all correlations were significant at the .001 level except for three out of 21 correlations reported (see Table 19, p. 91). On tenth-graders all intercorrelations were highly significant (see Table 20, p. 92), and the same finding was true on upper-elementary children (see Table 21, p. 92).

Data are reported on two validation trials (see Table 14, p. 84). These data seem out of place in a scoring manual. The .10 level used on the *elaboration* subtest seems to be straining at gnats, but perhaps tests on creativity can still be acceptable when their validity coefficients are significant at the .10 level or even the .20 level.

The interscorer reliabilities were based on two scorers in two tables and on four scorers in another table. Twenty-one subjects were used in one test-retest study. Similar studies at Cornell (Clark, 1964) did not find the .01 significance level that Yamamoto did. The high correlations reported in Table 15 (p. 85) should have made someone suspicious.

In their study of a sixth-grade population in California, Elnora Schmadel, et al., (1963) reported test reliabilities on the "Minnesota Tests"; *fluency*—.72; *flexibility*—.61; *originality*—.58; and *elaboration*—.75. They concluded their investigation by stating

that the validities of the test were scarcely high enough for use in individual counseling.

One of the basic assumptions in the scoring scheme which needs to be verified is that each of the four types of scores: *fluency*, *flexibility*, *originality*, and *elaboration* would be additive, because they are comparable. In the reviewers' opinions the area relationships have not been established, much less, linear relationships.

REFERENCES

- Clark, Nancy W. "A Study of the Relationship of Creativity, Preference for Open-Structure Learning Experiences, and Teacher-like for Pupils in Grades 3-6." Unpublished master's thesis, Cornell University, 1964.
- Schmadel, Elnora, Merrifield, Philip E., and Johnson, Henry S. "Use of the Minnesota Tests of Creative Thinking with a Sixth Grade Population." Paper presented at annual meeting of California Educational Research Association, Santa Barbara, California, 1963.
- Torrance, E. P. *Guiding Creative Talent*. Englewood Cliffs, New Jersey: Prentice-Hall, 1962.
- Torrance, E. P. "Problems of Highly Creative Children." In Walter B. Barbe (ed.) *Psychology and Education of the Gifted*. New York: Appleton-Century-Crofts, 1965.

R. C. PAXSON AND
R. W. BOYCE
Troy State College

Educational Testing for the Millions by Gene R. Hawes. New York: McGraw-Hill Co., 1964. Pp. xi + 290.

This book is not an angry cry against testing as is found in *They Shall Not Pass* or *The Tyranny of Testing*. Since testing like death and taxes is an inescapable event, the aim of the book is to explain to parents with a minimal of technical jargon the purpose and use of standardized tests. The author believes that objective tests have an important role to play in school. After reading the book, one feels that a parent should be in a position to use test data obtained from professional personnel like counselors, psychometrists, and psychologists in reflecting with the child possible courses of future vocational activity. Specifically, the book tries to help the parent to evaluate the chances that their child has of entering a college and of graduating. Since the purpose of the book is to *explain* and not to *criticize*, the role of test scores in college admissions is not debated. The author does review "clinical vs statistical" type of research in the area of college admission. It should be pointed out that it would take sophisticated parents to be able to use the book meaningfully.

Parents are introduced to standardized testing by a review of the 25 most commonly used tests at various levels of instruction. Each review contains information regarding the purpose and use of the test, grade level at which it is used, typical test content, and background on the publisher. Since the chapter is free of technical terms, the intent and tasks of each test should be readily grasped by the parents.

The next chapter contains an excellent review of intelligence testing. It discusses such matters as "What do intelligence tests measure?", "Is your child's IQ constant?", "Can you raise your child's IQ?", "What is the impact of culture on IQ?". In addition, technical terms like normal distribution, reliability, validity, standard error of measurement, and deviation IQ are discussed in non-technical language. Each term is explained by means of examples, diagrams, and charts.

The area of achievement testing is introduced by a review of "objective vs subjective" test-utility debate. The chapter discusses such issues as (a) influence of achievement testing on curriculum; (b) under-achieving and over-achieving; and, (c) the relationship between scores on achievement and intelligence tests. The areas of interest and personality testing are discussed in separate chapters. The purpose, use, value, and limitations of each type of test are reviewed. Again various controversial issues are presented. For example, Donald E. Super is used as a reference for supporting the use of interest tests, whereas John W. M. Rothney is cited to point out their limitations. Using quotations from two individuals who would be considered "experts" by their colleagues in exploring an issue is a typical device used by the author. The selection of quotations and "experts" is excellent in this respect.

The focal point of the book is a 70 page chapter entitled "Can your child get into college—and win a scholarship worth as much as \$10,000." The title of the chapter reflects the implied tone present in certain sections of the book that in order to sell books to the general public you must use emotionally loaded words, phrases, or ideas. Which is more effective, a section labeled simply "scholarship" or "win a scholarship worth as much as \$10,000?" Notwithstanding this difficulty, the chapter contains a comprehensive review of college admission procedures. The *Scholastic Aptitude Test* of the College Entrance Examination Board, the American College Test Program, selectivity of colleges, prediction of success in college, effect of coaching on admission tests, and the problem concerning who should go to college (anyone) are among the many topics discussed. Several expectancy tables are developed to show parents how they can use their child's test scores in estimating probable success in college. The author cautions the reader several times that the information obtained from an expectancy table is a

probability statement rather than a definitive statement. The ideas, observations, and examples contained in this chapter would have utility for most counselors and guidance workers.

The last two chapters discuss the nature of testing for professions, jobs, and armed forces as well as probable trends in standardized testing. The appendix contains a place to record the child's performance on various types of tests. It also reports by state the probable chances of securing test scores from a school.

It can be inferred from the above statements, that the information contained in the book is very similar to that found in many introductory textbooks on test and measurements. Appropriate "expert" opinion and research data are presented in conjunction with the exploration of various issues. Since the book is intended for the general public specific references are usually not given. In addition the author does use words such as eminent, prominent, and leading in introducing the "expert." Because the book minimizes the use of technical jargon and employs many charts and graphs, the parent after reading the book should benefit more from an interview with a counselor regarding their child's future activity than those who did not read the book.

It might even be advisable for those undergraduates who become confused with test and measurement procedures to look at this book for clarification of clouded issues. In summary, the volume is of value to the general public because it explains rather than condemns testing procedures. Although the author is favorably predisposed towards testing, he presents sufficient divergent information to allow parents to come to their own conclusion regarding the merits of testing.

HENRY KACZKOWSKI
University of Illinois

Introductory Statistics for the Behavioral Sciences by Robert K. Young and Donald J. Veldman. New York: Holt, Rinehart, and Winston, 1965. Pp. vii + 435.

Introductory Statistics for the Behavioral Sciences presents an innovatory and appealing approach to a one-semester course in introductory statistical inference. The most obvious departure from "conventional" texts in the field is the use of a programmed-learning technique to provide both acquisition and drill materials. A second, less obvious departure from tradition can be found in the selection of topics. Those included are said to reflect "the more stable developments in statistics during the past fifteen years."

These topics cover a rather startling range for a beginning course in statistics. After a brief introductory chapter which seeks to outline scales of measurement, there follows a chapter on graphs and distributions. Chapter 3 is a very valuable introduction, for the algebraically naive student, to single summation notation—an effort

paralleled later in the book (Chapter 12) for multiple summation notation. The first confrontation with "real" statistics starts on page 49 with a conventional sequence of chapters on measures of central tendency, variability, and standard scores. Chapter 7, titled "Foundations of Statistical Inference," is probably the weakest section of the book. It provides a very sketchy outline of the concept of probability—and in a somewhat breathless fashion tumbles simultaneously through illustrations of a number of terms which probably would have benefited from definition earlier in the book. In leisurely contrast, Chapter 8 is devoted entirely to a consideration of the notion of normal distributions and the area under the normal curve. Chapter 9 covers, very carefully, the standard error of the mean.

So far one might feel—and possibly rightly—that the authors' approach is scarcely new. However, the traditionalist would be surprised to find that, although the t distribution receives attention in Chapter 10, no attempt is made to present tests for the significance of differences between sample means by using t . Instead, greater attention is devoted (Chapter 11) to the F -test. This departure is justified on the grounds that ANOVA (analysis of variance) is more general and "easier to learn" than is the t test. One might feel, however, that by so arranging their material, the authors forego the chance to discuss the scientific method and hypothesis testing in general. Chapters 13 and 14 present one- and two-way ANOVA well and in a convincingly, simple fashion. Together, these chapters represent the high point of the book. The final two chapters, 15 and 16, return to more traditional topics for introductory courses in chi-square and in correlation and prediction.

The book is divided into two sections—text and programs. The text material is the result of the joint effort of both authors; the programs are the work of the first author alone.

The text sets out a brief overview of each of the 15 topics in 16 chapters. On the whole, the topics are clearly and simply presented in a style that should be attractive to the beginning student of statistics. Mathematical sophistication demanded of the student is very low; and in any instance where the authors consider difficulties might exist (such as in handling the summation notation), ample practice is provided in the programmed section to familiarize the student with the processes involved.

The avoidance of too great a demand on quantitative competence perhaps becomes a criticism when it is carried to the point where there are so few exemplifying problems worked in the text (and, therefore, minimal illustration of the use of these techniques in current research) and few problematic exercises at the end of each chapter. And, of those problems that are given, most require very little more computational work than do the steps in the programmed materials. To many who would use this book, the de-

emphasis on substantial quantitative material could, however, be seen as an advantage.

There are other occasions in the text when one feels a little uncomfortable. For example, the distinction between parameters and statistics seems to have been delayed unnecessarily long so that rather awkward "explanations" become necessary for defining the standard deviation with a denominator of $N - 1$. So, too, the symbolism adopted for the probabilities of multiple events looks strange. But these are minor sources of discomfort which should normally be readily dealt with in the course of instruction.

The authors claim greatest success with their approach in the laboratory, and there would seem little reason to doubt that realization of such success would be the case. There are numerous instances among the programmed items when other than the given answer could be acceptable. (Example: "If a distribution is skewed, it is not——" Given answer: "normal.") It would, therefore, appear to be advantageous to be in a position to discuss alternative responses with the student. Given the facilities to operate along the suggested instructional lines, the book would seem to provide a very suitable focus.

Although it is open to questions of ambiguity, the programmed material for the most part is quite good. Its purpose consistently appears to be to establish a minimum of vocabulary and an opportunity for very simple calculation. If for some students the approach seems rather tedious, for that group of students who suffer from an unwarranted "fear" of statistics, here lies the opportunity for considerable reward.

The content of the programs (all but two chapters are followed by at least 100 questions in the programmed format) not only supplements, but also adds to the text so that their careful, consistent working-through is a necessary part of reading the book. It is in this sense that the book is far more of an instructional aid than a reference source—a distinction prospective users should bear in mind. One of the very considerable advantages in the use of the programmed material is that it does make the student fairly independent of the instructor and free to work at his own pace.

The Appendix contains the eight statistical tables that are referred to in the text. The printing format of some of these leaves a little to be desired—but they will, no doubt, serve their purpose.

One further point might be noted. It is the omission of lists of reading materials to which the student can go to further his study of a topic. This is a particularly serious omission when one of the implied values of programmed instruction is that individuals can realize their different rates of learning. For those students who can finish a topic earlier than their colleagues (or who have a heightened curiosity), there is no indicated path to follow.

Overall, one cannot help but feel enthusiasm for this attempt to compromise between the traditional statistics text and the rather disjoint programmed texts. The programs are nicely balanced with textual material, are presented in a most ingenious fashion for a permanently-bound volume, and would seem to have considerable potential for achieving quite sophisticated goals with students who normally show some reluctance to tackle conventional statistics courses. The topics attempted are, to a degree, daring in their scope. Although none is presented in depth (and after all the book is a beginner's text), acquaintance with some of the techniques that have been traditionally regarded as more advanced should facilitate subsequent instructional courses. Provided that facilities are available for using the book as the authors suggest, in a laboratory-type setting, it should have a warm reception amongst those responsible for a first course in statistics.

PETER A. TAYLOR
University of Illinois

Intermediate Correlational Methods by Andrew R. Baggaley. New York: John Wiley & Sons, 1964. Pp. ii + 211. \$5.75.

This short introduction to correlational methods and factor analysis will undoubtedly receive two markedly different receptions. It will be read with mounting apprehension by many psychometricians and factor analysts and with mounting appreciation by most teachers. And that is largely as it should be, since this is the standard reaction to texts and since Baggaley has explicitly undertaken to write a textbook rather than a handbook. As a text the book has many advantages and also a serious shortcoming.

Baggaley's introductory chapter gives an exceptionally good short discussion of the distinction between experimental and correlational methods and of the strengths and weaknesses of each. It does make the common error of implying that factor analysis cannot be used as an experimental method even though writers such as Cattell have shown this viewpoint to be false. It might have been well to emphasize that experimentation and correlation are only two markers on the continuum of methods for obtaining and evaluating knowledge and that each technique at times shades indistinguishably into the other.

The second and third chapters introduce the Pearson product-moment correlation coefficient and its point-distribution alternatives. Baggaley does a workmanlike job of relating correlation to the concept of linear functions and linear prediction, and of illustrating the uses of the phi-coefficient and the other alternative forms of the Pearson r . These two chapters do not explicitly treat the concept of covariance, which is certainly of central importance and should not be left for the student to induce.

The theory of multiple regression and correlation is developed gradually, in easily digestible stages. The pivotal condensation method for simultaneous solution of a number of regression coefficients is employed. What the method lacks in elegance is more than made up for in ease of presentation. Rather amazingly, partial correlation is dismissed with two paragraphs "... because practitioners will not often deal with groups of persons completely homogeneous with respect to a particular continuous variable" (p. 86).

The general discussion of reliability and validity in Chapter 6 is quite orthodox and most of the views expressed are the standard ones. Baggaley does, however, obviously have grave doubts about the theoretical validity of construct validity, for he dances around and around the concept without ever quite meeting it face-to-face. The presentation of the Kuder-Richardson Formula 20 and its meaning and uses is probably as good as exists anywhere in the literature.

Chapter 7 deals with standard error of estimate, standard error of measurement, the Spearman-Brown formula, attenuation and a number of the more straightforward aspects of classical test score theory. The assumptions of each are clearly stated and most of their limitations commented upon. This chapter is excellent. It comes close to the ideal compromise between theoretical development and derivation on the one hand and intuitive appeal and practical application on the other.

The next chapters are less noteworthy. The reviewer's greatest doubts about the book arise from the sections on factor extraction and simple-structure transformations. Baggaley's working hypothesis is that most psychologists employ the centroid model and desk calculators to do their factor extractions and then use graphical rotation methods to find a simple structure. Following this assumption, he devotes almost the whole of this section to these two techniques and makes only the barest mention of principal axes and maximum likelihood methods of factor extraction and varimax and other objective criteria used in analytic rotations. A very few years ago this emphasis would have been quite appropriate, but a brief glance at the literature on these topics for the last two or three years tells a different story. It is increasingly apparent that the desk calculator and the protractor-and-square are rapidly becoming extinct in the factor analytic habitat and that the centroid model and graphical rotations are hard on their heels. Undoubtedly much more space and attention should have been given to the newer models and to consideration of computer techniques and the advantages they possess as well as to the limitations they impose. It seems unlikely that Baggaley's choice of material can be defended even on pedagogical grounds for the reason that much of this presentation is unclear. Many students who go through these chapters will find themselves in doubt as to what a centroid vector repre-

sents psychologically and also about the theoretical and computational differences between orthogonal and oblique rotations.

These criticisms fail to outweigh the favorable overall impression the book makes. There are several general features which add to this impression. Baggaley makes consistent good use of matrix concepts and notation in presenting his material. His problem sets and questions at the ends of the chapters are relevant and useful. Above all, Baggaley has a clear, levelheaded approach to the theoretical, computational, and practical aspects of correlational methods.

JAMES A. WALSH

Montana State University

Statistical Concepts: A Basic Program by Jimmy R. Amos, Foster Lloyd Brown, and Oscar G. Mink. New York: Harper & Row, 1965. Pp. ix + 125.

This new paper-bound programmed text presents an excellent introduction to the fundamental statistical concepts. The discussions cover all the basic concepts (e.g., frequency distributions, central tendency, variability, correlation, and inference) plus several of the more advanced topics (e.g., analysis-of-variance and regression). Very few formulas are listed; rather, the material is centered around intuitive discussions.

It is reported that the "typical beginning student" can complete the text in 5 hours. In the reviewer's opinion, these few hours would well be the best investment any beginning student can make.

Although the reviewer believes the major assets of this volume are found in the exceptionally clear expository writing style, he also commends the features of a combined index—diagnostic examination, the large easily-read type, and the clear and occasionally even humorous illustrations and figures.

This book is recommended as an excellent supplement for any course where it is desirable to build rapidly a statistical vocabulary and an intuitive understanding of concepts. Although this review is short, it is believed that this book could even be discussed in one short sentence—an excellent introduction for the verbally oriented undergraduate is afforded.

STEPHEN W. BROWN

University of Southern California

Who Goes Where to College? by Alexander W. Astin. Chicago: Science Research Associates, 1965. Pp. ix + 125. Hardback \$4.25, paperback \$2.25.

In admirable fashion Astin summarizes for the intelligent layman—especially the parent of a college-seeking high-school junior or senior—the chief procedures and findings of 16 of his publications and of studies by Holland, Nichols, and others. This bumper crop of National Merit Scholarship Corporation research was re-

ported in journal articles, mainly during the years 1961-65. Astin's readable monograph promptly makes it readily available.

Astin offers (on pages 57-83) a listing of five "freshmen input factors" and eight "environmental assessment" characteristics for each of 1,015 accredited four-year colleges and universities—nearly all such institutions in the United States. The rest of the monograph leads up to these 13 indices (each expressed in normalized standard-score form, with mean 50 and standard deviation 10) and carefully explains their origins and limitations.

The freshmen input factors are called intellectualism, estheticism, status, pragmatism, and masculinity. For example, the standard scores of entering freshmen (mostly men) at the highly selective California Institute of Technology (Cal Tech) on these are estimated to be, respectively, 81, 59, 60, 54, and 51, whereas those at Mills College (for women) also in California, are 60, 70, 63, 33, and 35. The scores at two closely related Massachusetts institutions, Harvard University and Radcliffe College, are, comparatively: 78, 76; 57, 72; 74, 69; 58, 34; and 66, 41. Thus Radcliffe freshmen are estimated to be 1.5 standard deviations (s.d.) more esthetic than Harvard freshmen, 2.4 s.d. less pragmatic, and 2.5 s.d. less "masculine." ("An entering student body with a high [masculinity] score would tend to have a high percentage of men, a high percentage of students seeking professional degrees (LL.B., M.D., D.D.S.), and a low percentage of students planning careers in social fields [such as education, nursing, social work, and social science]" (p. 55).)

Two of the eight institutional characteristics are selectivity ["Estimated Selectivity . . . is defined as the total number of highly able students who want to enroll at the college divided by the number of freshmen admitted" (p. 55)], and size, based on total full-time enrollment. For Cal Tech, Mills, Harvard, and Radcliffe these are scored 81, 63, 76, 81 for selectivity; and 51, 45, 69, 55 for size. Thus Cal Tech and Radcliffe are estimated to be equally attractive (81) to high-ability high-school juniors, and of the four institutions only Harvard is large (69).

The six "personal orientations" of the institutions, all based on the proportion of baccalaureates awarded by the institution in various fields, are realistic, scientific, social, conventional, enterprising, and artistic. The four institutions which have been considered were assigned the following scores on these.

	REA	SCI	SOC	CON	ENT	ART
Cal Tech	67	79	28	35	33	29
Harvard	47	64	40	45	65	55
Radcliffe	43	57	44	42	60	65
Mills	40	39	49	50	52	69

Cal Tech is 3.2 s.d. less "enterprising" than is Harvard—a find-

ing which means that the former had few (if any) graduates in 1960-61 who had majored in "such fields as advertising, business administration, history and political science (prelaw), journalism, international relations, and foreign service" (p. 56), whereas a large percentage of Harvard's graduates had majored in these areas.

Conversely, Cal Tech is 2 s.d. more "realistic" than Harvard, because engineering is a realistic field.

Radcliffe and Mills differ considerably only on the scientific orientation—an observation "based on the proportion of degrees awarded by the institution in various fields of natural science" (p. 56), with Radcliffe 1.8 s.d. ahead of Mills and 0.7 s.d. ahead of the national average.

To understand the 13 categories, one should read Astin's monograph carefully, because the one-word labels cannot themselves carry the entire weight of meaning. For example, "intellectualism" had the following factor loadings (principal-components analysis of 52 variables, with varimax rotation): *SAT-Mathematical*, [*SAT-Scholastic Aptitude Test* of the College Entrance Examination Board], .97; planning Ph.D. degree .84; *SAT-Verbal*, .78; scientific vocational choices, .75; "planning graduate work" and "National Merit scholars," .73 each; scientific major fields, .72; and "Placed in state science contest" and "median high-school grade average," .71 each.

Fifty of the 52 variables were secured from a questionnaire filled out by entering freshman in 248 selected colleges and universities, and the other two were *SAT-V* and *SAT-M* means for 43 of the 248 institutions. However, the freshmen input scores reported for the 1,015 institutions (including the 248) were *estimated* from institutional characteristics, rather than being ascertained directly. For example, the following eight predictors, with the standard partial regression coefficients as shown, were used to estimate intellectualism: estimated selectivity, .383; social orientation, —.329; conventional orientation, —.204; private nonsectarian, .189; Ph.D. output, .187; percent graduate students, .138; Catholic institution, .105; and Negro, .096. These produced a cross-validated multiple *R* of .85, when weights from 175 of 246 institutions were used on the other 71. One should note that this estimation procedure tends to increase the correlation between input factors and institutional-environment characteristics. Astin does not compute canonical correlation coefficients to check this *cross-dependence*, though on page 50 he compared the 10 intercorrelations of actual input scores for the 248 institutions from which questionnaire data came with the 10 intercorrelations of input scores *estimated* for all 1,015 institutions and found the latter to be larger in all but two instances.

Moreover, Astin does not mention the apparently great similarity of his measures to the six Allport-Vernon-Lindzey "Study of Values" evaluative attitudes.

The reader may wonder, too, why education of mother was not sought on the questionnaire, in view of the strong educational influences that mothers may have on their children. Education and occupation of father had high factor loadings for Factor III, Status, and moderate loadings for Intellectualism, Estheticism, and Leadership (later discarded).

Though finding the 13-score system highly suggestive for considering colleges that might be suitable for his high-school-senior daughter, this reviewer wished several times for separate norms for men and women. The so-called masculinity score, being a composite of percent males (standard partial regression coefficient, .657), Catholic institution (.267), enterprising orientation (.251), private nonsectarian (.152), Ph.D. output (— .145), and percent graduate students (.110), helps little to make sense out of a profile such as Pomona College's, particularly the 51 for "esthetic input" and the 55 for "artistic orientation," versus Scripps College's 72 and 73. Are the women at coeducational Pomona less esthetic and less likely to major in "such fields as fine arts, writing, languages, music, and speech" (p. 56) than are the women at all-female Scripps? How esthetic or enterprising are the *men* at Pomona? Presumably, it was not feasible for Astin to treat the men in coeducational colleges and universities separately from the women. When one considers the typically large sex differences on the "Study of Values" (men higher on theoretical, economic, and political, and women higher on esthetic, social—i.e., altruistic—and religious), he may suspect that sex differentiation would enhance the usefulness of Astin's otherwise valuable data.

However, as Astin pointed out to the reviewer in a private communication, the environment of a coeducational institution is probably the product or the interaction of various sex-differentiating characteristics, so that presenting data separately by sex may mislead the reader, unless he also is able to integrate the information in judging what the total environment is like.

Moreover, Astin indicated in the same communication, that, though complex universities are presented as single institutions in his book, data on which he is now working suggest strongly that the differences among colleges within a single university are much greater than the differences between universities.

In summary, the concerned parent (especially if he is a psychometrician!) and guidance counselor will want to purchase a copy of Astin's monograph to place alongside his other college aids for frequent use. The prospective college student's comfort and perhaps even his later success may be enhanced by using *Who Goes*

Where to College? Astin does not promise a great deal, however.

"Although colleges appear to differ in their effects on several aspects of the student's development, the size of these differential effects is small. In brief, the college actually attended by the student of high ability appears to make only a slight difference in his eventual career choice, academic and extracurricular achievements during college, academic ability, persistence in college, and the eventual level of education that he obtains. . . . Those characteristics of an institution that are generally believed to be educational assets—select student body, highly trained faculty, high faculty-student ratio, superior facilities (such as a large library)—appear to have little impact on the student outcomes that have been studied so far" (p. 89).

Researchers keenly aware of the markedly nonrandom assignment of students to colleges may suspect that certain "volunteering" variables not yet explored may well help counter the pessimism about college influence resulting thus far from the excellent institutional research of Holland, Astin, Nichols, and others. How, for instance, do persons identical with each other on the five input variables who choose quite different institutions compare on *other* variables relevant to later success? Astin has wise words for such researchers on pages 89–93. He has shown the path that leads beyond mere intra-institutional student accounting to the testing of important hypotheses.

JULIAN C. STANLEY
University of Wisconsin

An Introduction to Educational Research by Robert M. W. Travers. New York: The Macmillan Company, 1964. Pp. xxvi + 581.

The topic of scientific methods of research as applied to the behavioral sciences is one which is becoming increasingly emphasized in the training of graduate students in these areas. Murray Sidman's book, *Tactics of Scientific Research*, is an example of an approach to the problem in the field of psychology. In this book, his second edition, Travers addresses himself to the same types of questions from the standpoint of scientific research in education as Sidman considers in his book.

The initial discussion with which the book opens is that research is an activity with its own rules and terminology which apply to *all* scientifically studied fields. Examples are stated again and again to illustrate that the methods used by the historical "greats" of science e.g., Galileo, Mendel, and Faraday are the same methods as those which the educational scientist must adopt.

Travers discusses the use of theories, constructs, models, and different classifications of variables in research. His approach throughout the book is concerned with *what* the concept is, with what the use of the concepts contributes to the formulations of the

scientists, and with how concepts improve the accuracy and significance of scientific work.

The problem of choosing a problem which will result in fruitful research is one which Travers feels is important. He emphasizes that all scientists and graduate students in education must make this choice and that certain individuals seem more adept at identifying problems to study than are others. With great care, Travers points out not only why some problems would be more fruitful than others, but also how a possible research proposal may be evaluated.

There is a valuable chapter concerned not only with the kinds of variables studied in all sciences but also with their measurement—especially in education. Drawing upon his experience in working with graduate students in education Travers illustrates some of the more frequent problems of which the student must be aware.

Travers' strong bias that the most fruitful method of studying educational behavior is scientific is apparent throughout the book. Travers is dedicated to the training of research scientists who will be most likely to produce fruitful results from their study of important problems in education.

In directing his attention toward both the method of science and substantive problems in the field of education, Travers offers a real service to the person who wishes to conduct research in education. For the graduate student, the topics discussed are invaluable in building his ability to contribute to his profession. The concepts discussed are helpful not only in guiding the student in his first research project, but also in developing the student's skill in evaluating the research presented in the journals—a skill necessary for acquisition of knowledge in one's field of specialization.

In this introductory look at the problems of the science of behavior, the author has appropriately not attempted to delve into the philosophical depths of the questions he discusses, but has endeavored to address himself to a specific audience who are interested in educational research. It would appear that Travers has admirably accomplished his purpose in writing a text for the guidance of graduate students, of professors in schools of education, of teachers, and of all who are desirous of developing both an understanding of the methodology in educational research and some degree of skill in formulating and investigating significant educational problems.

REFERENCE

Sidman, Murray. *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*. New York: Basic Books, 1960.

GRETCHEN BRIEGLER TIMMERMANS
University of California
Santa Barbara

The Opposite-Form Procedure in Inventory Construction and Research by Jin Ong. New York: Vantage Press, 1965. Pp. 72. \$2.50.

This little monograph may help revive interest in the construction of opposite forms of self-report devices. Such forms, which date at least from 1923, have tended to be discredited recently on the basis of low correlations with the original forms.

Noting that Symonds in 1925 secured an r of .67 between opposite forms, each containing 115 items, for 102 subjects, Ong surveys various reversed items in order to determine why they usually fail to constitute near-comparable forms of the parent inventories. He does this comparison with considerable care and insight.

Then Ong devises seven opposite forms of published inventories, each containing 20 items. These, with a key, appear as the appendix of his book. His statistical results may have been influenced by the probable selective biases of his sample and, possibly, by the low motivation and fatigue of subjects faced with 140 items at each of two sittings, about one week apart; 61 of "over 120 adult university extension students . . . returned both forms completed" (pp. 44-45).

Ong concludes that if the deviser of opposite items applies certain rules, he can produce a form the scores of which will correlate well—perhaps even in the .90's for 100 items—with the original scores.

Though Ong warns that "generalization should not be over-extended to include all other devices such as achievement tests . . . [because] these may not have any opposites" (p. 58), the reviewer seems to recall that early in the history of group intelligence tests Otis devised parallel items for comparable forms by the device of changing an item such as "The opposite of love is _____" to "The opposite of hate is _____." Because this procedure tended to produce correlated errors from one form to the other, presumably it was abandoned in favor of more independent item content. Similarly, memory carryover from one form of self-report inventory to its opposite might produce a "spuriously" high correlation. Ong's rules and items should be of value for studying such memory effects as well as response biases.

JULIAN C. STANLEY
University of Wisconsin

The Structure of Knowledge and the Curriculum by G. W. Ford and Lawrence Pugno (Editors). Chicago: Rand McNally and Co., 1964. Pp. 105.

Underlying current proposals for change in American school curricula is the nature of the knowledge from which viable subject matter grows. As one reads history backwards—for example,

in the PSSC program—evidence is abundant concerning the analysis given to each basic discipline. Nevertheless, most schoolmen have only the vaguest notion of the structure and the operational propositions of the disciplines undergirding the school subjects. They have accepted the “new” mathematics, physics, English, and industrial arts on the hucksterish quality of the claims rather than on any solid understanding of whatever-it-is-that-is-“new.”

If this estimate of the contemporary scene is accurate, and if we assert its patent validity, two rather pessimistic predictions may be advanced. One, sustained curriculum reform efforts are imperiled. The American public and the schools, as a social institution, are unlikely to tolerate continued appeals for adoption of the “new.” Two, the older orthodoxy will be supplanted by a new rigidity possibly more restrictive and schismatic than that it replaced. Caught up in the excitement of this latter-day-progressive, reform movement, people welcome the “truth” as did the frontiersman in their brush arbor revivals. Diffusion of innovation has become a kind of secular evangelism heralded by its counterpart of “Have you been saved yet, brother?” and, not unlike the well-known town sinner who was “saved” at every tent meeting every summer, curriculum projects are being “adopted”—as soon as they are off the presses—without their being effectively understood. More importantly, their basic nature is seldom known, if even recognized as existing.

Possibly deterring these unfortunate consequences is the happy combination of (1) a general understanding, at least among some schoolmen, of the nature of the many knowledges and (2) the diligent extraction from this understanding of implications which may be instrumented into the school curriculum. Some murmurs and occasional rumbles have worked their way into the marketplace of ideas in the past five years; but like the “little” poetry magazines, published symposia and conference reports dealing with the nature of knowledge and with the curriculum are appropriately footnoted by the established orthodoxy and zealously guarded by those possessing the few extant copies. (Graduate students sign away possible graduation just to leaf through a copy!)

On this scene, a thin volume edited by G. W. Ford and Lawrence Pugno makes a significant and propitious debut. Published by a national house, the book is not fated to be footnoted *in absentia*. It will be welcomed, even embraced, by schoolmen who desire to know more of the relationship between knowledge and the school curriculum. Yet, *The Structure of Knowledge and the Curriculum* will not satisfy the present hunger. A future may well consider it more as a generalized stimulus in the curriculum field than as a particularly vigorous and ably argued thesis on the relationship of knowledge to the curriculum. For the insides of the

book do not live up to the title's billing; the volume seems blighted by a lack of proper concern for "the Curriculum," as it concentrates on interesting analyses of four disciplines. The several academic scholars—disciplinarians[?]¹—apparently are not expected to extract and to discuss implications from their analysis for the schools. Yet, someone must do it, particularly if the book is to have the desired value in influencing contemporary curriculum conversations. The editors might have done it; they did not. In such a book as this, a not unreasonable expectation is reference to some of the more impressive works and individuals from which and from whom the concern grew. Mentioned are only two previous projects—another university conference and a National Education Association project. Important as they are, these two projects are illustrative, but not exhaustive of the antecedents of the San Jose State College Curriculum Conference.

Probably most informative and productive of the essays are two by Joseph J. Schwab, a frequent and a major contributor to thinking about the nature of knowledge and the curriculum. The volume opens with his "Structure of the Disciplines: Meanings and Significances." Drawing on his intimate understandings in the sciences, Schwab discusses three crucial problems relating to the structure of disciplines: organization (What are they? How may they be known? How many are there?); syntax (patterns of inquiry and procedures for proof); and substantive structures (conditions of reliability and validity and situations of change). If only one essay in this volume is read, it should be this one. Certainly, it should precede attention to those that follow. Schwab's analysis—and it is only one man's notion—should add considerably to understanding of the basic issues. Too, implications for school programs emerge easily. Analysis of individual disciplines might proceed from the general framework he provides. Such a procedure would prove to be no easy task, but its products should serve to make available other competent analyses, and perhaps, to refine or elaborate his general framework.

Schwab's second essay, "The Structure of the Natural Sciences," primarily treats the syntax of the sciences. Discussing the syntax of stable inquiry, he provides excellent and stimulating criticism of the orthodox conception of scientific method exemplified by Dewey's famous five sequential steps. Schwab follows his cogently reasoned criticism with a reformulation of short-term scientific inquiry which, if adequately noted and instrumented, could have profound influence on teaching strategies and curriculum formulations. His discussion of fluid inquiry, or long-term syntax, and the quite varied substantive structures of science should prove both salutary and rewarding. The importance of his observation that the basic structures of the sciences *differ* is under-discussed. But the

point is advanced. Until it is adequately assessed, it will remain a disturbing and disquieting irritant to those who would misconceive the several sciences as having basically a common structure and common modes of inquiry.

T. H. Lange's essay on "The Structure of Mathematics" is, indeed, an erudite, interesting exposition which, if nothing else, presents to the neophyte and the untutored pedant a history of the change of a mathematician's thinking about mathematics. Unfortunately, as the discipline that before all others has been the cause of so much educational revolution, it fails most miserably to relate the mathematical milieu, as he would define it, to the curriculum. Mr. Lange states that mathematics can *only* be defined through active involvement and experience in mathematics—specifically in the creation and development of it. Alas the poor parents, teachers, and textbook writers! But, despite the handicap he gives the reader, Mr. Lange does provide an interesting contrast of "older" and "slightly newer" mathematics. To this point, he maintains a scholarly attitude; however he suddenly digs a hole, discards his scholarly appearances, and under the title of "much newer mathematical work" quotes verbatim, for three pages, a newspaper article from the *San Francisco Chronicle* on Paul Cohen's arithmetic, to no enhancement of Mr. Cohen. Lange ends nicely with a graded bibliography for the weak at heart.

Graham Wilson, on the other hand, blunders rather early in his "Structure of English," for as well read as he surely is, his rather simple dichotomy of English into language and literature puts the average thinking man, particularly the average thinking teacher, on a skyhooked library shelf supported with threads drawn from an excellent bibliography. In fact, the references are all that save this essay from total collapse. (He discusses English, not as language and literature, but as literature, grammar, composition, and literary criticism.) Of the domain of literature which one hopes would include all accepted literary art styles there is only vague reference. Critical theory dealing with the ways of knowing literature is treated, but it is not even suggested that this is what children might be learning. And since he reminds us that teachers know little of language, one wonders why he does not indict them for their misunderstanding of criticism. Here his failure to cite anyone since Cleanth Brooks—Northrop Frye for instance—might raise an eyebrow or two. Perhaps Mr. Wilson might be excused for not relating English, which he knows well, to the curriculum, which he apparently does not know well—let it be said that he did try—but there is another fault, more serious, because it is less likely to be noticed. That is his failure to mention children's literature and reading experiences or at least to delineate that he was not talking about them. Of course, it may be that these areas do not rightly

belong to "English." The real point is that Wilson has discussed components of what many people understand in English without confronting the question, "What is English?" Maybe there is no discipline of English, but rather, a field of studies.

Michael Scriven's "The Structure of the Social Studies" sidesteps any direct consideration of the distinctive elements of the nature of knowledge in the several social disciplines. His avoidance of what presumably is the central purpose of the set of essays does not, however, invalidate his contribution. He makes a case for the primacy of history, geography, and psychology in the theoretical and pedagogical structure of the social sciences. His position might have drawn more support had he delineated the fundamental nature of these disciplines. As it is, perhaps the most important position advanced embraces both the desirability of a *multi-disciplinary* approach in the social studies curriculum and the central importance of ethics in the social studies. On these points, there can be hearty applause. Still others must analyze the nature of the individual social disciplines if fundamental curriculum reform in the social studies is to be achieved.

This book has one major liability; it is simply a "non-book." It is a collection of essays—presented originally as lectures at what must have been an interesting conference—which are related to one another only in the sense that each presents an analysis of the basic structure of the discipline underlying one of the school subjects. Such collections are not to be decried; in fact, more should be encouraged. However, the conditions of synthesizing such a collection into a "book" demand features inadequately developed in this work. In the main, the introductory material provided by an editor should help illumine the collection as a whole as well as desirably each individual presentation. But unfortunately, in this book, the reader must slog through the individual chapters without the aid of any perspective, relationships or questions except his own.

Yet, the availability of this volume will help focus attention on the nature of knowledge and the curriculum. Graduate curriculum courses will probably use it as an important source. As it is widely used, however, it must not be allowed to preempt the field. Its availability could stifle the very discussion it can serve best to provoke. It presents only one point of view about four fields; for these positions to become dogma would be tragic. Instead, the volume should stimulate vigorous dissent, formulation, test, and dialogue. One must hear from other conceptualizations which also are valid. If these consequences are realized, this volume will have received high tribute indeed.

O. L. DAVIS, JR. AND
JOHN M. KEAN
Kent State University

Social Pressures in Informal Groups by Leon Festinger, Stanley Schachter, and Kurt Back. Stanford: Stanford University Press, 1963. Pp. x + 197. \$5.00.

After subtracting two chapters and adding two dollars, Stanford University Press has reissued *Social Pressures in Informal Groups*. This book, first published in 1950 by Harper and Brothers, is subtitled "A Study of Human Factors in Housing." Because of a modification in the usage of the term *human factors*, one might imagine today that the study pertained to the engineering of habitations to make them more adaptive to man. Instead, the book revolves around a study of the social relations among inhabitants of a temporary housing development immediately after World War II.

One justification for republishing the book at this time is the historic importance of the study as the progenitor of several sociopsychological studies of housing, including Festinger and Kelley's Regent Hill study and Back's more recent Puerto Rico study.

Social Pressures consists of nine chapters—seven chapters pertaining to a study of the social relations in a housing community, one chapter delineating a theory of group structure and group standards which emerged from the study, and one chapter describing the now familiar method of matrix multiplication in sociometry. The appendix contains an evaluation of the study procedures employed: informants, observations, interviewing, experimentation. An index rounds out the book.

How and why are groups formed? How does the group exert influence on its members? How do members resist group influence? In an attempt to answer these questions, Festinger, Schachter, and Back studied a housing project for married veteran students attending M.I.T. This community, named Westgate, consisted of 100 single-family houses arranged in nine courts. Six of the courts were identical in arrangement, with each consisting of 13 houses forming a U-shaped court. Half the houses (either 6 or 7) in each court were somewhat larger than the others and were reserved for families with children. As the houses were completed, they were occupied by applicants according to a first-come, first-served basis. A year later, in 1947, an adjacent housing development, Westgate West, was completed and was occupied by 170 veteran families. Westgate West consisted of 17 two-story converted Navy barracks, each holding 10 families. No experimental research was undertaken in Westgate West.

Data were collected in Westgate at two different times. The first set of data, collected in the Spring of 1946, a full year after Westgate had been occupied, consisted primarily of responses to three standardized open-ended questions put to each of the 100 wives by six trained female interviewers. These three questions were: (1) What three people in Westgate or Westgate West do you see most

of socially? (2) What do you think of the tenants' organization? (3) Are you active in the tenants' organization? The tenants' organization had initially a protective basis, which rapidly evolved into a social basis. The second set of data, collected a month later, consisted of responses of 90 families to the single question, "Have you heard of any publicity that the tenants' organization is getting?" This question was asked during the two evenings following the daytime planting of a rumor with two wives in the same court, and of a related rumor with two wives in another court.

On the basis of the responses to these four questions, as well as of the site layout, the authors define the terms physical distance, friendship choice, subgroup, group, group cohesiveness, group standard, group prestige, and communication channel. The authors state several conclusions: (1) the probability of a friendship choice is inversely related to physical distance, (2) group cohesiveness is inversely related to the number of subgroups (cliques) within a group, (3) group cohesiveness is directly related to the strength of group standards, (4) group cohesiveness is directly related to group prestige, and (5) friendship channels are directly related to channels of transmission of relevant information.

Even though the authors claim to have arrived at these conclusions based only on the responses to the four interview questions above, and on the site plan of the housing project, the conclusions are sometimes not consistent with the data. For instance, regarding the inverse relation between friendship choice and distance, the authors conclude that "the greatest number of choices were made to people living closest to the person choosing, and the choices decreased continuously as distance from the home of the chooser increased." However, the correlation between distance (front door-to-front door) and number of choices at each distance is .17 for Tolman Court and .10 for Howe Court, with both correlations in the direction opposite to that appropriate to the authors' conclusion. Had the greatest number of choices been made to people living closest to the person choosing, then one would expect frequent choices to be given to adjacent neighbors. Yet only one person in Tolman Court gave more than one of her three choices to her adjacent neighbors, and there was only one such person in Howe Court too. Results for the remaining courts could not be computed, since complete sociometric data were not given for these courts. Sociometric choices in the other courts may have followed a pattern consistent with the authors' conclusion; presenting these data once might have been more appropriate, and would have taken no more space, than presenting three times the identical Tolman-Howe data.

It is equally as difficult to understand how the authors, on the

basis of the data presented, could have come to the conclusion that "the existence of a friendship between two people also implies the existence of an active channel of communication." In this study friendship was measured by sociometric choice, communication channel by the transmission of the rumors planted simultaneously in Tolman Court and Howe Court. The rumors were transmitted 14 times, but only four of these transmissions were to friends. (The authors place this latter figure at six, clearly inconsistent with the sociometric data they present on pages 134, 139, and 145.) Again, the data appear to support a position just opposite to that espoused by the authors.

The reissue of *Social Pressures*, true even to the original typographical errors, will be welcomed again by those readers with a bent for formulations uninhibited by empirical evidence.

EDWARD LEVONIAN

University of California, Los Angeles

The Psychology of Learning Applied to Teaching by B. R. Bugelski. Indianapolis: The Bobbs-Merrill Co., Inc. Pp. xiv + 278.

Can human engineering be applied to education? Bugelski believes that it can and thereby increase the efficiency of the teacher. The haunting question "If a student fails to learn, is it the teacher's fault?" is answered in the affirmative by Bugelski. As a consequence, teachers should be knowledgeable in the fundamentals of learning so that they are constantly seeking "the better way" of doing things.

This is not a typical educational psychology textbook. It focuses only on the instructional aspect of the teacher's role. As Bugelski states, "We shall ignore the controversies and the broad systematic positions and concentrate on the occasional wisdom that appears in the writings of this or that theorist." However, his selection of theorists is one sided: they all belong to the behavioral school of psychology. Bugelski believes that by minimizing the theoretical aspect, the teacher will be more receptive to practical advice. As a result, experimental evidence, extended quotations, diagrams, and charts are used to supplement pertinent observations rather than as symbols of eruditeness.

Bugelski believes that the teacher is not interested in "nonsense" or in vague generalities, but in specific advice to commonly occurring classroom problems. In raising this question, he reviews the issues as to whether a psychologist should be primarily concerned with "pure science" or with the practical application of theory. The argument is reminiscent of a solution offered during World War II to counter the German submarine menace: boil the ocean! Should the classroom teacher be responsible for the discovery of the practical aspects of learning theory? Bugelski feels

that teachers need help in performing the necessary synthesis.

In five chapters Bugelski discusses the learning theories of Pavlov, Thorndike, Hull, Skinner, Tolman, Watson, Guthrie, and Mowrer. Although each chapter contains a thumbnail sketch of a theory, the review highlights only those aspects that have some pertinence to teaching. Bugelski's synthesis of each theory serves as an excellent base for the section that contains the practical implications of the theory. Additional chapters discuss the following topics: early and late training, attention, reinforcement, forgetting, transfer, programmed learning, and technical aids to education.

A teacher may question the utility of the "practical advice" found in this book. Thyne (1963) makes the following observation on this point,

A teaching-technique works because it fits. It must therefore be designed for the express purpose of making learning take place; the gears of teaching must be cut so as to engage the cogs of learning. Teaching-techniques cannot be assessed in a vacuum. They are tools; like keys they must be judged not by their intrinsic elegance, nor conformity to contemporary fashion, nor historical interest, but by whether they turn the locks of learning.

One may question whether the format used by Bugelski is the best for the purpose he has in mind. Instead of deducing "pearls of wisdom" from theoretical exploration, perhaps it could have been more convenient to arrange the material by tasks or typical concerns of teachers (i.e. habit formation, habit breaking, fostering skills, etc.). In all fairness to Bugelski he does have an excellent summary chapter in which he lists 58 suggestions that may enhance the teacher's efficiency in the classroom. His remarks about the use of examinations is of particular value to all teachers.

Although Bugelski states rather clearly that he is only concerned with the instructional aspect of the many roles played by the teacher, "practical advice" on the other tasks would make the book more complete. As Allport (1961) points out there are "matters of fact" and "matters of importance." In neglecting "matters of importance," the author deals with a fragment of the student rather than with the total student. As an aid in teaching "matters of fact" the book is most helpful. This criticism is based on the idea that educational experiences should transcend the mere inculcation of basic skills and knowledge. There are some who would disagree with this assumption. Allport (1961), reflecting on this point, said:

The biggest error most teachers make is to present to students their carefully drawn out conclusions when they themselves lack the raw experience from which these conclusions are fash-

ioned. . . . The teacher can best open channels of experience and by his obiter dicta sometimes lead students to see the value potential in the experience.

The above remarks are made in the spirit of exploring some of the issues inherent in the teaching act rather than merely criticizing the contents of the book. The premises on which a book is based should be explored. The writing is clear and lucid. The various examples given by the author help translate theory into action. Since most of the pertinent issues of behavioral learning theory are discussed, the book can be useful in courses for teachers where learning theory is discussed. Its use on the undergraduate level will depend on the orientation of the course. The main purpose of the book is to point to "better ways of teaching." For the most part, Bugelski has succeeded in doing this.

REFERENCES

- Allport, Gordon W. "Values and Our Youth" *Teachers College Record*, LXIII (1961), 211-219.
- Thyne, James. *The Psychology of Learning and Techniques of Teaching*. London: University of London Press, 1963.

HENRY KACZKOWSKI
University of Illinois

Boring, E. G., *History, Psychology, and Science: Selected Papers* by R. I. Watson and D. T. Campbell (Editors). New York: John Wiley & Sons, Inc., 1963. Pp. xxii + 372. \$8.95.

This book consists of a collection of papers written by Edwin G. Boring during the years 1921 through 1963. It enables the reader to follow Boring's thinking, on a remarkable variety of topics, throughout the years, and serves as a monument to this exceptional person, who helped mold psychology and provide it with form and substance.

The editors approached their task with imagination and dedication. It is all too easy to prepare a collection of papers by stringing them together in chronological order. This procedure requires no great plan; yet is generally safe from criticism. Happily, Watson and Campbell did not take this easy way out. They have organized the papers around some central themes, and have thereby produced a book, not just a collection of warmed-over reprints.

The volume is divided into five sections: *The Zeitgeist and the Psychology of Science*, *The History of Psychology*, *The Scientific Method*, *The Mind-Body Problem*, and *The Psychology of Communicating Science*. Within each section the papers are arranged logically around a central theme. For example, the first section starts with "Eponym as Placebo" which was Boring's address as

Honorary President of the Seventeenth International Congress of Psychology at Washington, D. C., on August 20, 1963. This is a provocative paper asking its readers to pay attention to the history of history—i.e., to man's account of what is happening and not simply to the events themselves. It also asks its readers to be aware of possible distortions to objectively caused by one's reverence for the "Great Men" of history.

The most substantive section, as one would expect, is "The History of Psychology," for Boring is best known as an historian of psychology. The selected essays, which deal with issues and personalities, provide, if not new insights, at least new appreciation of the role of psychophysics and of the influence of evolutionary theory and of measurement. The essays on scientific method and on the mind-body problem furnish an outlet for Boring's erudition as a philosopher as well as a psychologist. The final section consists of some essays on the art of book reviewing as well as thoughts concerning praise and criticism as adjuncts to communication. These papers originally appeared in "CP Speaks," a feature of *Contemporary Psychology*. Although pleasant, the papers do not have sufficient content to warrant being called "The Psychology of Communicating Science," which is the title of this last section.

Reading the selected papers is the next best thing to having a face-to-face conversation with Boring. In fact, because many of the papers were prepared as speeches, the book constitutes a one-sided conversation, in which the reader is really the listener. Needless to say Boring is never dull as a speaker.

HAROLD BORKO

System Development Corporation

The Education and Guidance of the Ablest by John C. Gowan and George D. Demos. Springfield, Ill.: Charles C. Thomas, 1964. Pp. xiii + 511.

The authors have provided a convenient compendium of current practices and writings on the gifted child. They have also been willing to express their own opinions. However, the paucity of sound research evidence to support what is being said and done in handling gifted children is disconcerting. Perhaps, the talents of the gifted should be sought to strengthen research on the gifted.

The unwary reader will encounter opinions expressed by the authors for which widely divergent views may be held. For example, after quoting a conclusion from page 362 of H. C. Hunt's *Practice of School Administration*, Gowan and Demos state: "This theory explains many things . . . why intelligence over the world seems to be on the rise, regardless of the differential fertility theory, why upper class children appear brighter, why first born

children appear brighter, and why bright adults seem to become brighter as they get older. It offers a great deal of hope for the wise use of the school curriculum." This quotation appears "non-sequitur" to the reviewer from Hunt's statements, as no further proof is provided for these sweeping statements.

The authors recommend that the "counselor should be permissive, intraceptive, client-centered and non-authoritarian." Evidence for these suggestions is not presented. Instead, in discussing the counseling of gifted girls, the authors indicate that the counselor should direct gifted girls to prepare for teaching as well as marriage.

No systematic effort is made in the book to caution unwary educators about the pitfalls of personality inventories. The use of paper and pencil personality tests is recommended for understanding underachievement. Although passing reference is made to two critical articles of this concept, the authors proceed to use the concept without really describing the objections that have been expressed by able psychometricians.

On page 293, the authors state: ". . . it will be generally found that the level indicated by the achievement test lags behind that of the intelligence test, and the marks and grade-level lag even farther behind the achievement test results." A single study by W. D. Lewis, published in 1944, seems to be the basis for this assertion.

Space does not permit citing other examples of statements or concepts that are presented in the book without sufficient qualifying remarks being attached. It is hoped that the reader will be sufficiently wary and will seek adequate data to determine the validity of the opinions that have been expressed. He will be handicapped when he seeks to check source works, as the bibliography is not directly referenced in the text. He will also find several citations missing from the bibliography.

Despite the limitations described above, the book does provide much useful information on the education of gifted children. Many different school programs are described; objectives, discussed; research agencies and programs, cited; and a comprehensive bibliography, presented. A complex view of giftedness is emphasized with attention focused on creativity rather than on convergent thinking. Chapter V on creativity provides an excellent summary of the work that has been done in the last few years in this area.

The educator interested in improving programs for the gifted will find succorance in this book. Educational researchers may be stimulated to conduct research to test the tenability of some of the categorical statements that are made.

WILLIAM COLEMAN
Coleman and Associates

*Passages from the "Idea Books" of Clark L. Hull. "Psychology of the Scientist: II. Clark L. Hull and His 'Idea Books'" by R. B. Ammons. "Psychology of the Scientist: III. Introduction to 'Passages from the "Idea Books" of Clark L. Hull'" by Ruth Hays. "Psychology of the Scientist: IV. Passages from the 'Idea Books' of Clark L. Hull" by Clark L. Hull. *Perceptual and Motor Skills Monograph Supplement*, 9-XV (1962), 800-882. \$3.00. Also distributed as articles in, and separates from, *Perceptual and Motor Skills*, XV (1962), 800-802, 803-806, 807-882.*

Scientific behavior tends to be identified with its end-product. This is especially true where the end-product approaches, or seeks to approach, formal theory and mathematical expression, as in Clark L. Hull's work in psychology. Yet, to understand scientific behavior, it is necessary to study not only its achievements, but also the human activity that is the creative scientific life.

Hull was unique in the hopes he held for psychology as a "natural science," in the degree to which he himself contributed to the fulfillment of these hopes, and in the documentation of these hopes in his idea notebooks. It is appropriate, therefore, that personal material scattered through Hull's 73 "idea books" be published as part of the series, "Psychology of the Scientist," of the journal, *Perceptual and Motor Skills*.

Two brief comments precede "Passages . . ." In the first, R. B. Ammons describes Hull's importance by reference to obituaries, journal citations, and ratings. He comments on the "intensely personal and private" nature of the writing and suggests questions for further study. In the second, Ruth Hays, Hull's secretary and technician for 22 years and selector of the excerpts, reviews the significance Hull attached to his "idea books" and lists some of their contents.

"Passages . . ." begins with an October, 1902, entry entitled "My Diary." At 18 years of age Hull wrote: "... I may some day rise to a position among men where I can look back with pride at the time when I was an obscure chore boy" (p. 807). The last entry is dated February 29, 1952, about 10 weeks before Hull's death. He wrote: "... false notions in physics from time to time have been cleared away at once by certain insights but this has not happened in psychology or behavior. Possibly something like that will be done by my 'System'" (p. 882). In between, most entries show Hull planning and evaluating his work and himself. Professional and personal goals and methods of goal achievement are set forth in relation to skills, limitations, ingenuities, age, health, machines, scientific controversy, stimulation of others, stimulation by others, publication, and the needs of psychology. On March 5, 1916, he decided (p. 814) to "... be a pure psychologist . . .," choose "... a

limited task . . .," and ". . . become the supreme authority . . ." in "*. . . the psychology of abstraction and concept-formation. . .*" Hull's doctoral dissertation was on concept learning, but his major book-length accomplishments were on aptitude testing, hypnosis, and systematic behavior theory.

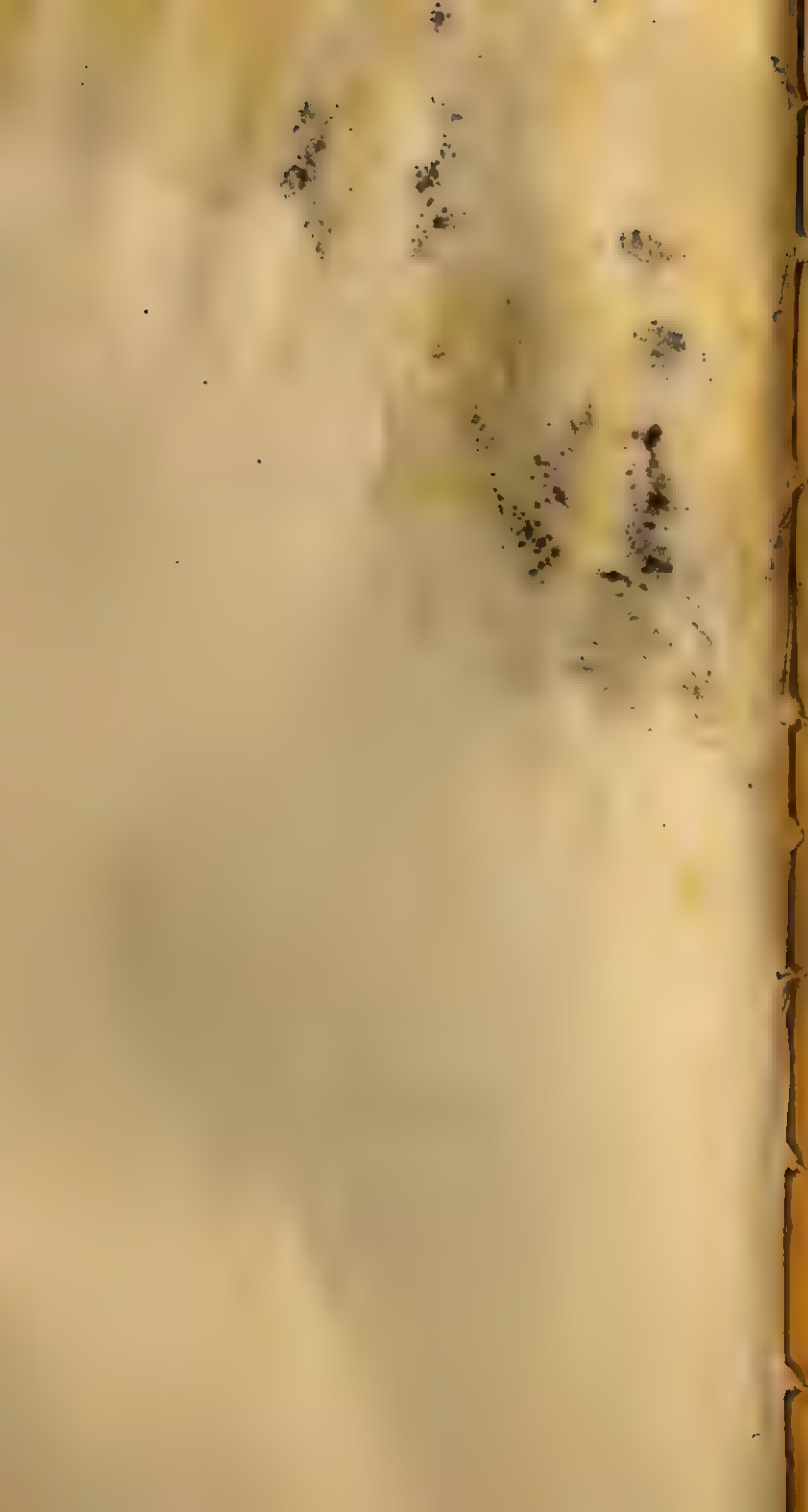
The personal and emotionally gripping nature of "Passages . . ." is such as that it can be read and enjoyed as literature and as supplementary reading in history and systems courses at the advanced undergraduate and graduate levels. The young researcher may find in "Passages . . ." useful tips on planning, organizing, and publishing experimental and theoretical work. The student of scientific creativity will want to raise at least two general questions: (a) Which of Hull's personality traits were necessary and which were epiphenomenal? and (b) How much of Hull's planning and self-evaluation came prior to overt successful behavior and how much after, and how much from others and how much from himself? Specific hypotheses bearing on one or the other of these questions ought to be testable in research on children, adolescents, and adults.

The great part of Hull's writing was, of course, objective and scientific. In reading "Passages . . ." one must recall that the excerpts were selected for publication because they were personal. The title, "Personal passages . . .," rather than "Passages . . .," would thus have been more accurate. The reviewer participated as an undergraduate in Hull's seminar, and served as his research assistant, in 1942-43. He found "Passages . . ." fascinating reading and not unlike Hull's overt behavior. Hull was fond of making asides in his seminar and in his conferences. The asides hinted at, and sometimes expressed, the emotional intensity contained in "Passages . . ."

Hull strove consciously to achieve greatness for himself and for psychology. He sought to solve, not to argue, relevant problems raised by the behavioral scientist, philosopher, and metaphysician. He spoke boldly but always without absolute certainty. He dared his seminar students to believe in, and work toward, the quantifiability of an individual human being's totality of behavior, even if ". . . the equation stretched to the moon." Perhaps "Passages . . ." will inspire some "boy" to comparable zeal, effort, and achievement.

WALTER C. STANLEY
National Institute of Mental Health







-1 DEC 1905



EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

- A Conjunction of Rank Order Typal Analysis and
Item Selection.* LOUIS L. MCQUITT 949
- A New Working Formula for the Split-Half Relia-
bility Model.* NAMBURY S. RAJU AND ISALAH
GUTTMAN 963
- The Use of Simulated Stimuli and the "JAN" Tech-
nique to Capture and Cluster the Policies of
Raters.* JAMES C. NAYLOR AND ROBERT J. WHERRY,
SR. 967
- Number of Scale Points and the Reliability of Scales.*
S. S. KOMORITA AND WILLIAM K. GRAHAM 987
- Development and Classification of Models for Multi-
variate Analysis.* MARY C. REGAN 997
- Reliability of Composite Ratings.* JOHN E. OVERALL 1011
- A Generalization of the Median Test.* DALE E.
GUTTMAN 1023
- The Use of Inference as a Research Tool.* CLIFFORD
C. COUBSON 1029

<i>A Factor-Analytic Study of Spontaneous-Flexibility Measures.</i> FRANK B. MAY AND ALAN W. METCALF	1039
<i>Experience, Expertness, and Ideal Teaching Relationships.</i> WILLARD E. REITZ, PHILLIP S. VERY, AND GEORGE M. GUTHRIE	1051
<i>Evaluative Responses to Affectively Positive and Negative Facial Photographs: Factor Structure and Construct Validity.</i> CARROLL E. IZARD AND JUM C. NUNNALLY	1061
<i>A Validation of Howard's Test of Change-Seeking Behavior.</i> GEORGE DOMINO	1073
VALIDITY STUDIES SECTION	1079
BOOK REVIEWS	1165
INDEX FOR VOLUME XXV	iii

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Contributors receive one hundred reprints of their articles without charge. Manuscripts should be sent in duplicate to G. Frederic Kuder, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia.

Subscription rate, \$10.00 a year, domestic and foreign. Single copies, \$2.50. Back volumes: Volume V or later, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

Editor: G. Frederic Kuder

Associate Editor: W. Scott Gehman

Managing Editor: Geraldine R. Thomas

BOARD OF COOPERATING EDITORS

WILLIAM V. CLEMANS

Science Research Associates, Inc.

LOUIS D. COHEN

University of Florida

HAROLD A. EDGERTON

Performance Research, Incorporated

MAX D. ENGELHART

Chicago City Junior Colleges

E. B. GREENE

Chrysler Corporation

J. P. GUILFORD

University of Southern California

JOHN A. HORNADAY

Houghton Mifflin Company

E. F. LINDQUIST

State University of Iowa

FREDERIC M. LORD

Educational Testing Service

ARDIE LUBIN

U. S. Naval Hospital, San Diego

SAMUEL MESSICK

Educational Testing Service

WILLIAM B. MICHAEL

*University of California,
Santa Barbara*

HOWARD G. MILLER

*North Carolina State University
at Raleigh*

P. J. RULON

Harvard University

C. L. SHARTLE

Ohio State University

KENDON SMITH

*The University of North Carolina
at Greensboro*

THELMA G. THURSTONE

*University of North Carolina
at Chapel Hill*

HERBERT A. TOOPS

Ohio State University

JOHN E. WILLIAMS

Wake Forest College

E. G. WILLIAMSON

University of Minnesota

DOROTHY ADKINS WOOD

*University of North Carolina
at Chapel Hill*

VOLUME TWENTY-FIVE, NUMBER FOUR, WINTER, 1965



A CONJUNCTION OF RANK ORDER TYPAL ANALYSIS AND ITEM SELECTION*

LOUIS L. McQUITTY
Michigan State University

LINEAR models for the measurement of individual differences are often very exacting in their requirements; they specify the model as linear and data are required to conform to it. Consequently the model is helpful in the selection of test items. Many techniques of item analyses have been developed for selecting sets of items which conform to linear models.

Pattern-analytic methods need analogous techniques to assist them in the selection of items; they need techniques which will select items that conform to a pattern-analytic model.

If a pattern-analytic model is to be effective in the selection of items, it is helpful to have it precise with respect to the attributes which operate in the selection process. This principle should not, however, conflict with another desirable characteristic of pattern-analytic methods. Pattern-analytic methods should not require the model to specify the kinds of interrelationships reflected by the data; they should not, for example, superimpose linear relationships when the predominant patterns are either curvilinear or disjunctive.

A Review of Rank Order Typal Analysis

Rank Order Typal Analysis (McQuitty, 1963) meets the above requirements and can be elaborated so that it can be used in the fashion there outlined.

* Appreciation is expressed to Miss Margaret Thomas for the analysis of data.

Rank Order Typal Analysis defines a type as a category of people of such a nature that everyone in the category is more like everyone else in the category than he is like anyone in any other category.

Symbolic representation of two categories of people, one fulfilling and the other one failing to fulfill the above definition of a type, are shown in Figures 1 and 2 respectively.

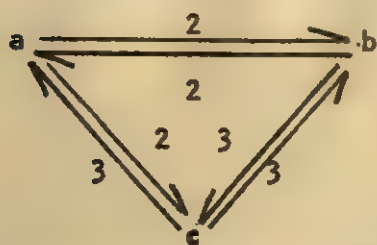


Figure 1. A Typal Category of People

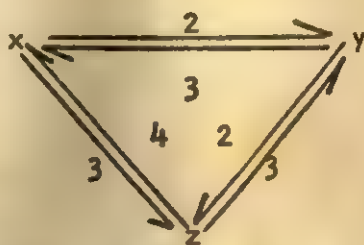


Figure 2. A Non-typal Category of People

In the case of Figure 1, Person *a* has *a* most like himself, *b* second most like him and *c* third most like him. A similar condition holds for each Person *b* and *c*. Consequently, each person in the category is more like everyone else in the category than he is like anyone not in the category; there is no rank in the category higher than the number of people in the category. When this latter condition is satisfied the definition of a type is fulfilled.

Figure 2 reflects an exception to the above kind of internal consistency; Person *x* has *z* fourth most like himself, *y* third most like him, and he is first most like himself. This means that there is some other person in some other category who is second most like *x*. As a consequence, Figure 2 contains one rank which is higher than the number of persons and does not therefore qualify as a type; whenever a category contains one or more persons with a rank higher than the number of persons in the category, it fails to qualify as a type.

It is a simple matter to examine a matrix of interassociations between people in such a fashion as to isolate all of the categories which qualify as types. The task is accomplished by examining the matrix serially one column at a time.

The first step is to take a matrix such as the one shown in Table 1 and convert it into a matrix of ranks as reported in Table 2 to show for example that the person most like *x* is *x* himself and then *w*, *v*, and *z* in that order.

TABLE 1

Hypothetical Associations between People

	<i>w</i>	<i>x</i>	<i>y</i>	<i>z</i>
<i>w</i>	71	60	65	35
<i>x</i>	60	72	50	40
<i>y</i>	65	50	73	45
<i>z</i>	35	40	45	74

TABLE 2

Rank Order within the Columns From Table 1

	<i>w</i>	<i>x</i>	<i>y</i>	<i>z</i>
<i>w</i>	1	2	2	4
<i>x</i>	3	1	3	3
<i>y</i>	2	3	1	2
<i>z</i>	4	4	4	1

If *x* forms a type with any one person of Table 2, it must be with the Person *w* who is second most like him (*x* being most like himself). In order to examine this possibility, we form a submatrix of *x* with *w*, Table 3, using the ranks from Table 2.

TABLE 3

A Submatrix from Table 2

	<i>w</i>	<i>x</i>
<i>w</i>	1	2
<i>x</i>	3	1

TABLE 4

A Submatrix from Table 2

	<i>w</i>	<i>x</i>	<i>y</i>
<i>w</i>	1	2	2
<i>x</i>	3	1	3
<i>y</i>	2	3	1

Person *x* does not form a type with Person *w*; the submatrix contains a rank larger than the number of persons in the matrix. Since Person *x* does not form a type of two persons with the one person second most like *x* (*x* being most like himself), Person *x* does not form a type of two persons with any other one person of Table 2.

If *x* forms a type with any two persons of Table 2, it must be with the persons second and third most like *x*. These are *w* and *y*. A submatrix of *w*, *x*, and *y* is formed, Table 4. It contains no rank larger than 3. It proves, therefore, that *x* forms a type of three people with *w* and *y* and furthermore forms no other type of three people from those of Table 2.

The analysis proceeds in this fashion until Column *x* is completed. In case of a 4×4 matrix, as in Table 2, a column has been completed when three variables have been included in a submatrix, as they have in Table 4; the analysis for *x* was completed in the above operations.

In the general case, Column *i* of an $n \times n$ matrix has been completed when a submatrix of $n - 1$ variables has derived from a study of Column *i*. If all of the variables of a matrix were included in a submatrix, they would by necessity meet the above criterion but

the criterion would no longer operate as a test of the presence of a type.

Other columns of the matrix are analyzed in a similar fashion. In order to save time, steps should be taken to avoid duplications in the types isolated. For example, if Person x forms a type of three persons with Persons w and y , then each of these latter persons will yield the same type; they, therefore, need not be examined for this possibility.

In case of ties of more than two indices, it is helpful to follow a method of assigning ranks which deviates from the usual method. In the new method, if Rank 7, for example, is followed by three tied indices such as 23, 23, 23, and then a larger index, say of 25, the three tied indices are assigned a rank of 10—the highest of the ranks in the series required for the three tied entries. The next entry is assigned a rank of 11.

In the general case, tied indices are each assigned the highest rank of the series required to cover the tied values. This approach prevents, for example, three tied cases from first forming three types of two cases each and then a type of three cases when all three cases are equally alike.

Once the data have been analyzed in the fashion outlined above, all possible types have been isolated, i.e., all possible types which derive from all of the data, with every item equal in weight to every other item; other types can possibly be isolated in terms of reduced sets of data as will now be shown.

Multiple Rank Order Typal Analysis for the Isolation of Intersecting Types

The Method

Rank Order Typal Analysis can be elaborated both (a) to select items which are internally consistent in the description of types, and (b) to extend each of several initial types into a comprehensive typology.

The latter operation can yield a comprehensive system of intersecting typologies. The comprehensive and intersecting typologies are assumed to be due to a complex of pressures acting on people (institutions or other objects). Each major pressure is presumed to impel toward the development of a typology. However, each typol-

ogy is hidden because several pressures are active. As a result the typologies intersect in their development to such an extent that they become obscure.

A method is needed to untangle and display the several typologies expressed in the behavior of people (institutions or other objects). Multiple Rank Order Typal Analysis for the Isolation of Intersecting Types was developed for this purpose.

Assume that a matrix of interassociations has been analyzed and that it yields the following types: $T_1, T_2, T_3 \dots T_n$. For each T , determine the items on which all subjects of the type agree in the answers they give. Call these items sets, $S_1, S_2, S_3 \dots S_n$, corresponding to $T_1, T_2, T_3 \dots T_n$. For each S , compute a matrix of interassociation using all of the subjects but only the items of the set. This will give matrices $M_1, M_2, M_3 \dots M_n$, corresponding to $T_1, T_2, T_3 \dots T_n$. Perform a Rank Order Typal Analysis on each $M_1, M_2, M_3 \dots M_n$. Each M might yield types so that we might have (a) for M_1 : $T_1^1, T_1^2, T_1^3 \dots T_1^{n_1}$; (b) for M_2 : $T_2^1, T_2^2, T_2^3 \dots T_2^{n_2}$; (c) for M_3 : $T_3^1, T_3^2, T_3^3 \dots T_3^{n_3}$, etc. Each of these T 's could be analyzed to yield corresponding sets of items and matrices which could in turn be pattern analyzed. This kind of continuation would in many situations seem to be inappropriate. A purpose of the approach is to eliminate noise (items which are interfering with the complete emergence of types). Hopefully this would generally be accomplished without analysis beyond that of matrices $M_1, M_2, M_3 \dots M_n$.

Rank Order Typal Analysis possesses two characteristics which render it particularly helpful in the selection of items. It utilizes a stringent definition of a type; in order for a category to qualify as a type, every individual in the category must be more like every other individual in that category than he is like any individual in any other category.

The definition is applied in such a fashion that its degree of stringency varies with the validity of the data to which it is applied. As the data become more valid (either through application of the method or through some other effective method of item selection) the application of the definition generally becomes more stringent.

When a relatively high proportion of items is relatively invalid for the isolation of a type, then all of the members of that type are not usually isolated by Rank Order Typal Analysis during the first application; only three out of a total of ten members might, for

example, form the first type, initially. The definition of a type is not stringent when applied to only three of ten actual members. This application of Rank Order Typal Analysis would nevertheless eliminate some invalid items. As a result, a reapplication of the method, using the reduced set of items, would probably admit more members to the type. In order that an item could then be retained, it would have to be characteristic of more members; the stringency of the approach would have been increased.

If Rank Order Typal Analysis is applied to any matrix of inter-associations between persons (institutions or other objects), it will yield one or more types. It will always yield at least one type. This is because the highest entry in a matrix is always reciprocal; i is highest with j and j is highest with i ; they form a type by the above definition of a type.

If only one type, as opposed to several, is portrayed by the data, then there is a great possibility of the outcome being due to chance alone. If the outcome is due to chance alone, then the successive stages of the extended analysis will be relatively unsuccessful in yielding types.

Some Characteristics of the Method

Multiple Rank Order Typal Analysis for the Isolation of Intersecting Types has a unique feature in terms of the types it isolates; it searches out initial types and then examines each type to see how many additional types can be defined in terms of the items characteristic of each initial type. Additional people are categorized into additional types at the expense usually of decreases in the number of characteristics which the members of the new types have in common.

A union, for example, might have been developed in several companies of the construction industry and had impact on unions developed later in other companies. Simultaneously a union might also have developed in companies of the trucking industry and also had impact on some of the unions developed later in other companies.

In a case like this the first application of Rank Order Typal Analysis could be used to help assess the joint influence of the construction and trucking industries in the typology of these and other companies.

The method might first classify together into types (a) companies

of the construction industry, and (b) companies of the trucking industry, as well as other companies into other types.

The extended method could then be applied. Using only the items common to the companies of the construction type, the results might reflect some influences of the construction union in the typologies of the companies of the other unions. Using, on the other hand, only the items common to the companies of the trucking type, the results might reflect some influences of the trucking union on the unions of the other companies.

Depending on the purpose of the study, it might be desirable to use in the one case only items common and unique to the companies of the construction type (as distinct from the trucking type) and in the other only those common and unique to the companies of the trucking type (as distinct from the construction type).

Each company might have several classifications. Furthermore, each of several types of companies might have several classifications into larger types. These results can all be portrayed, and the investigator can make his choice of those he wishes to retain on the basis of any criteria which he can show to be appropriate for his purpose.

An approach of the above kind assumes that types are evolutionary and interwoven with one another; they are presumed to be dynamic, reflecting at the same time both formative and disintegrative forces which are at times in conflict, and the investigator can be appropriately concerned with both (a) disentangling the types, and (b) selecting out the various combinations of types of concern to him, depending on his scientific interests and purpose as expressed in his study.

An Application of Rank Order Typal Analysis

In order to illustrate Rank Order Typal Analysis, it was applied to indices of interassociations among eight industrial companies in terms of their union-management relations (McQuitty, 1954). The interassociations are shown in Table 5. The interassociations are reported in terms of agreement scores. The companies were assessed on 32 continuous variables involving union-management characteristics. These were dichotomized and each company was either above or below the median in terms of each variable. Two companies agree on a variable if they are both either above or below the

TABLE 5
*Agreement Scores between Companies**

	A	B	C	D	E	F	G	H
A	32	29	16	16	14	6	11	7
B	29	32	17	17	13	6	8	10
C	16	17	32	26	10	8	9	13
D	16	17	26	32	10	12	11	11
E	14	13	10	10	32	21	17	13
F	6	6	8	12	21	32	19	17
G	11	8	9	11	17	19	32	24
H	7	10	13	11	13	17	24	32

* Data from McQuitty, 1954

median but not if one is above and the other is below the median. The agreement score between two companies is the number of variables on which the companies agree.

Companies A and B are in the construction industry, C and D—trucking, E—grain processing, F—metal products, and G and H—garment manufacturing.

Results of the Rank Order Typal Analysis

The Rank Order Typal Analysis of Table 5 yielded Types AB, CD, ABCD, EF, and GH as reported elsewhere (McQuitty, 1963).

An Application of Multiple Rank Order Typal Analysis for the Isolation of Intersecting Types

In order to illustrate Multiple Rank Order Typal Analysis for the Isolation of Intersecting Types it was applied to the above results from Rank Order Typal Analysis. We first determined the items on which the companies of the first type, Type AB, agreed. There were 29 such items as reported in Table 5. Using only these

TABLE 6
Agreement Scores between the Companies on the 29 Items of Type AB

	A	B	C	D	E	F	G	H
A	29	29	15	15	12	5	8	7
B	29	29	15	15	12	5	8	7
C	15	15	29	25	8	8	8	11
D	15	15	25	29	10	10	10	9
E	12	12	8	10	29	20	15	12
F	5	5	8	10	20	29	17	16
G	8	8	8	10	15	17	29	24
H	7	7	11	9	12	16	24	29

items, a new matrix of agreement scores between all of the companies was computed, as shown in Table 6.

In a similar fashion a matrix was computed for the items of each of the other four types, CD, ABCD, EF, and GH, and they are shown in Tables 7, 8, 9, and 10.

TABLE 7

Agreement Scores between the Companies on the 26 Items of Type CD

	A	B	C	D	E	F	G	H
A	26	25	13	13	10	5	8	6
B	25	26	14	14	9	4	7	7
C	13	14	26	26	7	7	7	9
D	13	14	26	26	7	7	7	9
E	10	9	7	7	26	19	14	12
F	5	4	7	7	19	26	14	14
G	8	7	7	7	14	14	26	22
H	6	7	9	9	12	14	22	26

TABLE 8

Agreement Scores between the Companies on the 13 Items of Type ABCD

	A	B	C	D	E	F	G	H
A	13	13	13	13	2	0	1	1
B	13	13	13	13	2	0	1	1
C	13	13	13	13	2	0	1	1
D	13	13	13	13	2	0	1	1
E	2	2	2	2	13	11	10	10
F	0	0	0	0	11	13	12	12
G	1	1	1	1	10	12	13	13
H	1	1	1	1	10	12	13	13

TABLE 9

Agreement Scores between the Companies on the 21 Items of Type EF

	A	B	C	D	E	F	G	H
A	21	20	12	12	5	5	7	6
B	20	21	13	13	4	4	6	7
C	12	13	21	19	4	4	3	7
D	12	13	19	21	6	6	6	5
E	5	4	4	6	21	21	13	10
F	5	4	4	6	21	21	13	10
G	7	6	3	6	13	13	21	18
H	6	7	7	5	10	10	18	21

TABLE 10

Agreement Scores between the Companies on the 24 Items of Type GH

	A	B	C	D	E	F	G	H
A	24	24	14	14	8	4	5	5
B	24	24	14	14	8	4	5	5
C	14	14	24	22	8	5	7	7
D	14	14	22	24	8	5	7	7
E	8	8	8	8	24	18	11	11
F	4	4	5	5	18	24	14	14
G	5	5	7	7	11	14	24	24
H	5	5	7	7	11	14	24	24

TABLE 11

Agreement Scores between the Companies on the 14 Items of Type FGH

	A	B	C	D	E	F	G	H
A	14	14	13	13	4	0	0	0
B	14	14	13	13	4	0	0	0
C	13	13	14	12	3	1	1	1
D	13	13	12	14	2	1	1	1
E	4	4	3	2	14	10	10	10
F	0	0	1	1	10	14	14	14
G	0	0	1	1	10	14	14	14
H	0	0	1	1	10	14	14	14

TABLE 12

Agreement Scores between the Companies on the 10 Items of Type EFGH

	A	B	C	D	E	F	G	H
A	10	10	10	10	0	0	0	0
B	10	10	10	10	0	0	0	0
C	10	10	10	10	0	0	0	0
D	10	10	10	10	0	0	0	0
E	0	0	0	0	10	10	10	10
F	0	0	0	0	10	10	10	10
G	0	0	0	0	10	10	10	10
H	0	0	0	0	10	10	10	10

Rank Order Typal Analysis was applied separately to each of these tables.

The Rank Order Typal Analyses of Tables 6 to 10, inclusive, yielded in addition to types already isolated, two new Types FGH and EFGH. The matrices for the items common to the members of each of these two types are shown in Tables 11 and 12. Each of them was analyzed by Rank Order Typal Analysis.

Results

Table 5 shows the original indices of association between the companies. Tables 6 to 12, inclusively, show analogous indices where only items common to the members of a type were used in computing the interassociations. A simple observation reveals that the latter tables reflect types more clearly than the original table, thus indicating that the method of item selection is effective in picking out the items indicative of types.

The results from Multiple Rank Order Typal Analysis for the Isolation of Intersecting Types are shown in Table 13. Starting with the items common to the two companies of the construction type,

TABLE 13
Results from Extended Rank Order Typal Analysis

Using Items of	Number of Items	Types Obtained and the Number of Items on which Members Agree						
Type AB	29	AB-29	CD-25	ABCD-13	EF-20	GH-24		
Type CD	26	AB-25	CD-26	ABCD-13	EF-19	GH-22		EFGH-10
Type EF	21	AB-20	CD-19	ABCD-10	EF-21	GH-18		EFGH-10
Type GH	24	AB-24	CD-22	ABCD-13	EF-18	GH-24		EFGH-10
Type ABCD	13			ABCD-13		GH-13	FGH-12	EFGH-10
Type FGH				ABCD-12			FGH-14	EFGH-10
Type EFGH				ABCD-10				EFGH-10

Companies AB, the reapplication of Rank Order Typal Analysis yielded Types AB (construction), CD (trucking), EF (grain processing and metal products), GH (garments), and ABCD (construction and trucking). Each of the other three diadic starts (Types CD, EF, and GH) yielded not only these same types (as did starting with Type AB) but in addition they produced Type EFGH (grain processing, metal products, and garments). Starting with Type ABCD failed (necessarily) to yield Types AB and CD, but it did agree with the others in yielding Type GH, but failed to yield Type EF. Instead it produced Type FGH and then yielded Type EFGH.

Starting with Type FGH yielded Types ABCD, FGH and EFGH, and starting with Type EFGH produced Types ABCD and EFGH.

The appearance of Type EFGH in six analyses but not in the first one is related to earlier findings in other studies. When the data were forced into types (without evaluation by an index of internal

consistency), Type EFGH invariably appeared even though all items were used (McQuitty, 1954 and 1960, for example). On the other hand when a criterion of internal consistency (as in Rank Order Typal Analysis) was applied to the members of a proposed type, Companies EF and GH did not qualify as a type in terms of all of the items of the original study (McQuitty, 1963).

In other words, Type EFGH appeared as an internally-consistent type only when some of the items were selected out of the original set.

Interpretation

The results illustrate two points: (a) The items used determine, of course, the types. (b) There are various meaningful points of departure for the selection of items in a typological study. One point of departure, for example, is to use items which have shown themselves to be relevant in the development of union-management relationships in general. Another point of departure is to use items thought from some other perspectives to be representative of a particular industry, such as construction or trucking. Each approach can yield one or more characteristic typologies in other companies. The methods of this paper can assist in disentangling and displaying the intermingled typologies.

In a more general sense, various sets of pressures act on institutions (people and other objects) of our world. Each of them provides forces toward the development of types. As a consequence, typologies are often intermingled and need, for purposes of study, to be separated out in some meaningful fashion, such as illustrated by the methods of this paper.

Summary

The method here developed grew out of the assumption that different sets of pressures operate on people (and institutions) as they develop. As a consequence, people (and institutions) develop various patterns of characteristics, or types, which become intermingled and relatively hidden from "unaided vision."

The method applies a stringent definition of a type, which enables the techniques to select out one or more initial types. These initial types are then used to select the items which both (a) portray these types more clearly, and (b) extend the several typologies to include

other persons (or institutions) which did not appear in the initial types.

The result is the possible disentanglement of overlapping and relatively hidden types; they are displayed as several relatively complete typologies.

REFERENCES

- McQuitty, L. L. "Rank Order Typal Analysis." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIII (1963), 55-61.
- McQuitty, L. L. "Pattern-Analysis: A Statistical Method for the Study of Types." in W. E. Chalmer, M. K. Chandler, L. L. McQuitty, R. Stagner, D. E. Wray, and M. Derber (eds.), *Labor-Management Relations in Illini City, Volume II*. Champaign, Illinois: Institute of Labor and Industrial Relations, University of Illinois, 1954.
- McQUITT, L. L. "Hierarchical Syndrome Analysis." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XX (1960), 293-304.

A NEW WORKING FORMULA FOR THE SPLIT-HALF RELIABILITY MODEL

NAMBURY S. RAJU AND ISAIAH GUTTMAN

Science Research Associates, Inc.

THERE are two major working formulas for computing split-half reliability coefficients—the Spearman-Brown formula and the series of formulas (e.g., those of Flanagan, 1937; Guttman, 1945; Mosier, 1941; Rulon, 1939) equivalent to

$$r_{ss} = \frac{4r_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + 2r_{12}\sigma_1\sigma_2}, \quad (1)$$

which has been attributed to Flanagan. The literature (e.g., Cronbach, 1951; Gulliksen, 1950) generally regards the Spearman-Brown formula as assuming $\sigma_1 = \sigma_2$ while the Flanagan formula does not.

The purpose of this paper is to present a new working formula for the split-half model. The formula is

$$r_{ss} = \frac{r_{12}(\sigma_1 + \sigma_2)^2}{\sigma_2^2}, \quad (2)$$

and has the following characteristics:

1. does not assume $\sigma_1 = \sigma_2$
2. is always equal to or less than the Spearman-Brown and equal to or greater than the Flanagan estimates
3. is a lower bound to the true reliability, except in one case where the situation is indeterminate.

Derivation

Let the obtained score for any individual be his true score plus an error score. Assuming that error scores are randomly distributed, we may define reliability as

$$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2}, \quad (3)$$

where σ_t^2 is the true variance. Splitting the test into two halves, then

$$r_{xx} = \frac{\sigma_{t_1}^2 + \sigma_{t_2}^2 + 2r_{t_1 t_2} \sigma_{t_1} \sigma_{t_2}}{\sigma_x^2}, \quad (4)$$

$$= \frac{r_{11}\sigma_1^2 + r_{22}\sigma_2^2 + 2r_{12}\sigma_1\sigma_2}{\sigma_x^2}. \quad (5)$$

Using r_{12} as an estimate of r_{11} and r_{22} (i.e., $r_{12} = r_{11} = r_{22}$), equation (5) may be written as

$$r_{xx} = \frac{r_{12}\sigma_1^2 + r_{12}\sigma_2^2 + 2r_{12}\sigma_1\sigma_2}{\sigma_x^2}, \quad (6)$$

and

$$r_{xx} = \frac{r_{12}(\sigma_1 + \sigma_2)^2}{\sigma_x^2}. \quad (7)$$

Equation (7) is a split-half estimate of the test reliability when the variances of and correlation between the halves are known, and applied to the situation where essentially all items are attempted. There is no assumption that $\sigma_1 = \sigma_2$. It is equivalent to a formula derived by Cureton (1958) who, however, did not seem to regard it as a working split-half reliability formula.

Relationship to Other Split-Half Reliability Formulas

The relationship between the three split-half reliability formulas is indicated by equation (8), assuming $r_{12} \geq 0$.

$$\frac{4r_{12}\sigma_1\sigma_2}{\sigma_x^2} \leq \frac{r_{12}(\sigma_1 + \sigma_2)^2}{\sigma_x^2} \leq \frac{2r_{12}}{1 + r_{12}} \quad (8)$$

The left inequality of equation (8) will be examined first. To show it is true, we need only show that equation (9) is true.

$$4r_{12}\sigma_1\sigma_2 \leq r_{12}(\sigma_1 + \sigma_2)^2 \quad (9)$$

Expanding equation (9) yields

$$2r_{12}\sigma_1\sigma_2 \leq r_{12}(\sigma_1^2 + \sigma_2^2). \quad (10)$$

Since $(\sigma_1 - \sigma_2)^2 \geq 0$, we have $2\sigma_1\sigma_2 \leq \sigma_1^2 + \sigma_2^2$. Multiplying both sides by r_{12} yields equation (10), which is identical to equation (9).

The right inequality of equation (8) will now be examined. To show it is true, rewrite equation (7) as

$$\frac{r_{12}(\sigma_1 + \sigma_2)^2}{\sigma_z^2} = \frac{2r_{12}(\sigma_1 + \sigma_2)^2}{(1 + r_{12})(\sigma_1 + \sigma_2)^2 + (1 - r_{12})(\sigma_1 - \sigma_2)^2} \quad (11)$$

Since $(1 - r_{12})(\sigma_1 - \sigma_2)^2 \geq 0$, we have

$$\frac{2r_{12}(\sigma_1 + \sigma_2)^2}{(1 + r_{12})(\sigma_1 + \sigma_2)^2 + (1 - r_{12})(\sigma_1 - \sigma_2)^2} \leq \frac{2r_{12}(\sigma_1 + \sigma_2)^2}{(1 + r_{12})(\sigma_1 + \sigma_2)^2}, \quad (12)$$

which is equivalent to the right inequality of equation (8).

Equation (7) as a Lower-Bound to the True Reliability

This section will examine conditions under which equation (7) will be less than or equal to equation (5), and where the relationship is indeterminate. The relationship is primarily a function of the relative sizes of r_{11} , r_{22} , and r_{12} since they are involved in the major assumption in the derivation of equation (7). It should be remembered that under the assumption that $r_{11} = r_{22} = r_{12}$, equations (7) and (5) are equivalent. Two conditions involving the relationship of r_{11} , r_{22} , and r_{12} may be distinguished (the special case $r_{22} = r_{12} < r_{11}$ is common to both):

- 1) $r_{12} \leq r_{11}$ and $r_{12} \leq r_{22}$
- 2) $r_{22} < r_{11}$ and $r_{22} \leq r_{12} \leq r_{11}$.

We assume that r_{11} , r_{22} , and r_{12} are all greater than zero and have the following relationship: $r_{12} \leq \sqrt{r_{11}r_{22}}$.

To show that equation (7) does not overestimate equation (5), the following relationship should hold:

$$r_{12}\sigma_1^2 + r_{12}\sigma_2^2 \leq r_{11}\sigma_1^2 + r_{22}\sigma_2^2. \quad (13)$$

A. Case I. $r_{12} \leq r_{11}$ and $r_{12} \leq r_{22}$

For any given σ_1 and σ_2 equation (12) will hold.

B. Case II. $r_{22} < r_{11}$ and $r_{22} \leq r_{12} \leq r_{11}$

Equation (7) will not be a lower-bound under all conditions of Case II. For $\sigma_2 \leq \sigma_1$, it will be a lower-bound, but for $\sigma_1 < \sigma_2$, the situation is indeterminate.

Using the algebraic principle that the arithmetic mean is equal to or greater than the geometric mean, we have

$$2r_{12} \leq 2\sqrt{r_{11}r_{22}} \leq r_{11} + r_{22}, \quad (14)$$

and

$$r_{12} - r_{22} \leq r_{11} - r_{12}. \quad (15)$$

If $\sigma_2 \leq \sigma_1$, then

$$r_{12}\sigma_2^2 - r_{22}\sigma_2^2 \leq r_{11}\sigma_1^2 - r_{12}\sigma_1^2, \quad (16)$$

which is equivalent to quotation (13).

If $\sigma_1 < \sigma_2$, equation (16) will not necessarily hold, and we cannot say that equation (7) will necessarily be a lower-bound.

Relationship to Kuder-Richardson Item Statistic Formulas

Guttman (1945) interpreted his split-half formula (L_4) as a two item special case of L_3 , which is identical to $K-R$ (20). Similarly, equation (7) of this paper can be interpreted as a two-item special case of $K-R$ (14) (Kuder and Richardson, 1937).

$$K-R(14) = \frac{\sigma_z^2 - \sum \sigma_i^2}{(\sum \sigma_i)^2 - \sum \sigma_i^2} \cdot \frac{(\sum \sigma_i)^2}{\sigma_z^2}, \quad (17)$$

where σ_i refers to item standard deviations. If the test consists of two items (or two half tests), this can be rewritten as

$$K-R(14) = \frac{\sigma_z^2 - (\sigma_1^2 + \sigma_2^2)}{(\sigma_1 + \sigma_2)^2 - (\sigma_1^2 + \sigma_2^2)} \cdot \frac{(\sigma_1 + \sigma_2)^2}{\sigma_z^2}. \quad (18)$$

However,

$$\frac{\sigma_z^2 - (\sigma_1^2 + \sigma_2^2)}{(\sigma_1 + \sigma_2)^2 - (\sigma_1^2 + \sigma_2^2)} = r_{12}. \quad (19)$$

Substituting equation (19) into equation (18) yields equation (7), showing the algebraic identity.

Summary

1. Equation (7) is an estimate of the split-half reliability of a test, assuming $r_{11} = r_{22} = r_{12}$, but not assuming $\sigma_1 = \sigma_2$.

2. Reliability estimates from equation (7) will always be equal to or lie between estimates obtained from the Spearman-Brown formula and estimates from the Flanagan type formulas which do not assume $\sigma_1 = \sigma_2$.

3. Reliability estimates from equation (7) are always lower bounds to the true reliabilities except in the case where $r_{22} < r_{11}$,

$r_{22} \leq r_{12} \leq r_{11}$, and $\sigma_1 < \sigma_2$. In this case, the situation is indeterminate.

REFERENCES

1. Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, XXVI (1951), 297-334.
2. Cureton, E. E. "The Definition and Estimation of Test Reliability." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XVIII (1958), 715-738.
3. Flanagan, J. C. "A Proposed Procedure for Increasing the Efficiency of Objective Tests." *Journal of Educational Psychology*, XXVII (1937), 17-21.
4. Gulliksen, H. *Theory of Mental Tests*. New York: Wiley, 1950.
5. Guttman, L. "A Basis for Analyzing Test-Retest Reliability." *Psychometrika*, X (1945), 255-282.
6. Kuder, G. F. and Richardson, M. W. "The Theory of the Estimation of Test Reliability." *Psychometrika*, II (1937), 151-160.
7. Mosier, C. I. "A Short Cut in the Estimation of Split-Halves Coefficients." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, I (1941), 407-408.
8. Rulon, P. J. "A Simplified Procedure for Determining the Reliability of a Test by Split-Halves." *Harvard Educational Review*, IX (1939), 99-103.

THE USE OF SIMULATED STIMULI AND THE "JAN" TECHNIQUE TO CAPTURE AND CLUSTER THE POLICIES OF RATERS*

JAMES C. NAYLOR AND ROBERT J. WHERRY, SR.

The Ohio State University

THE policy of a rater has been defined as "what raters do" when they are asked to respond to a series of stimuli (Naylor and Wherry, 1965). Further, one can define the process of "capturing" the policy of a rater as the extent to which one can predict the actions of that rater from the known characteristics of the stimuli he is being required to evaluate. For example, if we ask k judges to rank a group of n individuals with regard to general job competence, the way in which they are ordered by each judge would represent that judge's policy. To the extent that two judges did not order people in the same fashion their policies would be considered to be different. Also, the extent to which one could predict the rank order for a given judge using information about the individuals being ranked would represent the degree to which that judge's policy has been captured.

The process of capturing policies and of determining their general similarities and differences are two distinctly separate problems and have received considerably different emphasis on the literature. Generally speaking, almost nothing has been done concerning policy capturing, while a fair amount of research is available pertaining to clustering raters in terms of the similarities of their judgment. The intercorrelation and factor analysis of this type of data is the most

* This research was supported in part by the Air Force Systems Command, Contract AF41(609)-1596. The opinions expressed in this paper are those of the authors and not necessarily those of the Air Force.

typical solution. Indeed, almost all the research pertaining to the profile clustering problem is directly relevant (e.g., see Sawrey, Keller, and Conger, 1960; Cronbach and Gleser, 1953; Haggard, et al., 1959; McQuitty, 1957, 1960 and Nunnally, 1962).

Recently a new approach to both the policy capturing problem and the policy clustering problem has been proposed by Bottenberg and Christal (1961). Labelled the JAN technique (Judgment Analysis) by Christal (1963), the method groups raters in terms of the homogeneity of their prediction equations and is a special application of a still more general grouping model developed by Ward (1961). The JAN technique requires, however, that a reasonable number of stimuli be used in order to provide some stability to the individual rater equations and to guard against overfitting if a large number of stimulus dimension are employed. This places a severe restriction on the use of JAN in those cases where the stimuli are expensive to obtain and to measure. This is particularly the case when people are to serve as stimuli, since in order to use JAN it would be necessary to obtain a measure on each of the n persons serving as stimuli for each of the p traits assumed to be potential predictors of policy. Madden (1963), in the only reported use of JAN to date, overcame this problem of pre-measurement of stimuli by creating artificial or "simulated" stimuli. However, Madden only had 50 stimuli with 10 dimensions each, and he made no attempt to concern himself particularly with the relationships between predictors (the covariance matrix) which was created by assigning simulated scores to his stimuli dimensions. This matrix may indeed be important, particularly in those situations where it is necessary for the dimension scores to have a certain degree of relationship to maintain "face-validity" of the stimuli. For example, if the simulated stimuli represent people who have been measured on p different traits, it would be most unusual for certain traits not to be related to each other. Unless the scores on the stimuli reflect this relationship it is likely that the rater may (a) find it difficult to judge certain stimuli because they appear unrealistic and (b) become suspicious that the stimuli he is being asked to evaluate may not be "real" stimuli.

Fortunately, the problem of generating large numbers of scores on each of a number of dimensions which have a specified covariance matrix is not difficult as Wherry, Naylor, Wherry and Fallis (1965)

outline a procedure to accomplish this using the basic factor analytic model,

$$Z_{ij} = r_{i1}Z_{i1} + r_{i2}Z_{i2} + \cdots + r_{iq}Z_{iq} + \sqrt{1 - r_{i..}^2}Z_{iu} \quad (1)$$

where Z_{ij} is the predicted standard score received by individual i on dimension j ; the Z_{iq} 's represent the standard scores of individual i on each of the independent factors obtained from the correlation matrix; and the r_{iq} 's represent the factor loadings of dimension j with each of the q factors. The final term represents the product of the correlation of dimension j with a uniqueness factor, comprising the specific of dimension j and random error with the standard score Z_{iu} of individual i on a unique factor for each dimension j . In order to generate scores on p dimensions which will possess a given covariance matrix one need only to specify the underlying factor structure of the dimensions. One of the major purposes of this study was to evaluate the use of simulated scores generated with the Wherry et al. procedure as a means of providing stimuli for raters to evaluate. If large numbers of stimuli can be generated having specified covariance matrices which can be used satisfactorily with raters then one of the major restrictions of JAN is overcome.

A second question associated with JAN concerns the F test used as a criterion of determining the number of different policies which exist among the raters. This test, outlined in detail in Bottenberg and Christal (1961), evaluates the hypothesis that the two equations being grouped possess the same set of regression weights.¹ However, it is used sequentially and in a fashion such that each succeeding ${}_sR^2 - {}_{s+1}R^2$ must be larger than its predecessor, and the question of whether this test actually corresponds to the F distribution might be raised. An additional purpose of the research was to examine how well the significance test actually did accomplish the task of providing information concerning the number of different strategies or policies represented among a set of raters.

$${}^1F = \frac{{}_sR^2 - {}_{s+1}R^2}{1 - {}_sR^2} \times \frac{N - (k - s)P}{P} \quad (2)$$

- where: ${}_sR^2$ = maximum predictive efficiency of group at iteration s
 ${}_{s+1}R^2$ = maximum predictive efficiency at iteration $s + 1$
 k = total number of equations to be grouped
 s = iteration
 N = number of observations
 P = number of predictors

Method

Subjects

Fifty Air Force supervisors served as raters in the study.² All raters were E-9's in the specialty of jet aircraft maintenance (the 9 level in the Air Force is officially supervisory in nature). Raters were obtained from seven separate Air Force Bases with the n at each base varying from 1 to 26. Raters were selected at each base strictly on their availability for testing rather than by any random process.

Stimuli (Profile) Traits

Two hundred fifty profiles were produced using the Ohio State Correlated Score Generation Method (Wherry, Naylor, Wherry and Fallis, 1965). Each profile presented a score for that "man" on each of the 23 different traits. The scores ranged from one (very poor) to nine (very good) using the basic stanine system. The 250 scores for each trait were normally distributed, and the intercorrelation of profile trait scores differed only by sampling error from a "population" correlation matrix obtained directly from the basic factor structure used to generate the profile scores.

The basic factor structure used to generate the trait scores is given in Table 1, and Table 2 shows both the population correlation matrix between traits created directly from the factor structure (above diagonal) and the empirical correlations between traits obtained by intercorrelating the generated trait scores on the 250 profiles (below diagonal). Diagonal entries are communalities. The factor structure used to generate trait scores was one arbitrarily arrived at through the consensus of five "experts," all of whom were senior staff members on the research project. This F Matrix contained nine factors as being appropriate to describe the underlying dimensions necessary to explain the traits. These were (a) a general factor, (b) a compliance factor, (c) a motivation factor, (d) a leadership factor

² The study was actually carried out using four different Air Force Specialties, each with 50 raters. Since the data in all four cases was similar—as were the experimental outcomes—data from only one specialty is described in this article.

with two sub-generals, consideration and structuring, and (e) an intelligence factor with two sub-generals, knowledge and skill.

The twenty-three traits used on the profiles were selected in a pilot study as being the traits most likely to be relevant to job suc-

TABLE 1
Final Overall Trait Factor Structure (Theoretical)
Used to Generate Rates Profiles

Variable # (Trait)	General	Compliance	Motivation	Leadership	Consideration	Structure	Intelligence	Knowledge	Skill	h^2	Used on profiles
1	30	40	50				30			59	1
2	30	70								58	
3	40			40	60					68	2
4	20			40	50	30				54	3
5	40		40	20	30	30	20			58	
6	30		30				40	30	30	52	4
7	40		30	40	70					90	5
8	30			50	40	50	20			79	6
9	30		60			30	30	30	30	81	7
10	30	50					30		30	52	8
11	40	20	20	40		60				76	9
12	40			40		30	40	50		82	10
13	20	40					20			24	
14a	40	30		40	40		30			66	
14b	40	30		40	40		30			66	
15	30	30	30	50	40	30				77	11
16	40		30	45		70				94	12
17	40	20	60							56	13
18	30	20	30	40		40				54	14
19	20			50	60					65	15
20	30	20	60				30			58	16
21	30		40							25	
22	40		50				40		50	82	
23	40	60	30							61	17
24	30			40		40	40	40	20	77	18
25	40	50								41	
26	30	20	20				50	60	20	82	19
27	40						50	40	50	82	20
28	30	50	20				20		30	51	21
29	30						40	50	40	66	
30a	30	40	20			20				33	
30b	30	40	20			40				45	22
31	30						60	60		81	23
32	20						50	20	50	58	
33	30	40	20			20	20		20	41	
34	30						30	50	20	47	

TABLE 2

Theoretical Correlations (Above Diagonal), Empirical Correlations (Below Diagonal) and Communalities (Diagonal) between Traits Used to Describe Simulated Job Incumbents

	Variable Number																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	59	12	06	36	27	15	48	38	30	24	36	27	50	32	06	56	51	21	42	27	45	35
2	12	68	54	12	74	56	12	12	32	32	56	34	16	28	64	12	16	28	12	16	12	12
3	04	51	54	06	59	61	15	06	42	33	55	47	08	34	54	06	08	34	06	08	06	18
4	39	18	05	52	21	17	57	30	18	43	18	21	30	18	06	39	21	43	59	59	32	15
5	28	75	55	28	90	60	30	12	38	32	69	43	34	37	70	30	25	30	18	16	18	18
6	12	49	61	22	59	79	30	15	62	55	65	70	12	49	55	15	12	37	19	22	13	29
7	39	06	10	51	30	26	81	27	42	48	36	51	48	39	06	54	30	51	60	54	36	33
8	33	10	10	35	13	12	28	52	22	24	24	12	22	19	06	28	42	27	40	42	49	29
9	31	34	47	25	47	69	39	19	76	50	62	82	32	62	28	28	24	52	20	16	26	48
10	32	34	33	42	39	53	40	25	55	82	41	55	16	40	08	24	16	76	62	56	20	24
11	36	55	52	29	72	66	38	24	63	50	77	65	36	56	55	33	39	41	21	12	30	39
12	32	29	45	29	48	67	50	11	79	55	65	94	34	67	31	30	25	58	18	16	18	46
13	50	17	06	33	38	12	52	21	32	15	37	42	56	34	08	52	46	12	28	16	34	32
14	30	30	36	29	42	46	43	16	57	46	56	64	31	54	26	31	33	41	19	12	25	39
15	12	68	51	12	70	50	02	11	34	34	57	30	14	31	65	06	08	26	06	08	06	06
16	58	11	13	36	34	16	51	25	29	23	37	34	57	33	10	58	42	21	40	27	37	29
17	51	17	12	22	25	16	26	40	39	29	42	28	40	33	18	32	61	12	30	16	48	42
18	26	26	34	43	38	58	48	27	59	77	52	61	14	49	26	20	19	77	57	71	23	25
19	43	14	02	57	23	16	55	38	24	60	32	21	27	28	08	37	32	55	82	71	39	21
20	26	18	09	56	21	22	48	43	22	51	20	18	13	19	15	23	22	54	63	82	37	12
21	46	15	11	40	22	08	38	46	24	22	32	22	38	30	15	34	45	26	39	40	51	33
22	27	04	24	16	21	31	28	35	46	28	40	47	28	29	12	21	40	33	25	15	29	45
23	29	08	01	40	08	12	29	18	16	62	11	08	04	10	02	21	14	51	70	57	20	04

cess in that specialty. Figure 1 shows a sample profile. It is important to note that these profiles represented "men" in the specialty immediately below that of the raters themselves (i.e., the level they directly supervised)—in this case E-7's.

Procedure

The profiles were produced by a computer and high speed printer on 7" × 11" IBM paper as shown in Figure 1. Verbal anchor points were used at 5 points along the 9-point stanine scale. At the bottom of each profile was an additional scale—also on a nine point continuum—that was used by the rater to indicate his global criterion judgment about that profile. For any given rater all 250 profiles had the traits listed in the same order. However, 25 different (random) orders of traits were used. Since there were 50 raters, this meant that each order was presented twice, i.e., two raters were assigned to each of the 25 random trait orders.

Trait No.	Trait	Profile 1 Very good								
		Very poor	Poor		Avg		Good		Very good	
		1	2	3	4	5	6	7	8	9
15	Utilizes opinions of subordinates	5				
9	Maintaining discipline	.	.	.	4					
10	Ability to communicate	6			
23	Knows theory of aircraft systems	5				
18	Ability to teach others	7		
19	Knows tech orders-manuals-sops	6			
6	Willing to delegate work	6			
22	Careful with property and equipment	.	2							
13	Perseverance on the job	.	2							
21	Careful and thorough in inspections	5				
20	Proficient in trouble shooting	6			
12	Monitors work of subordinates	5				
8	Safe work habits	.	.	3						
1	Willingness to learn	.	.	.	4					
11	Willingness to train his men	5				
7	Plans and organizes his work	5				
5	Supporting his men	.	.	.	4					
16	Ability to perform under pressure	6			
14	Willingness to make decisions	8	
3	Informs his men	6			
2	Fairness and impartiality	.	.	.	4					
4	Job ingenuity	5				
17	Punctuality and dependability	.	2							
		1	2	3	4	5	6	7	8	9

Figure 1. Sample profile of a simulated job incumbent.

Raters were tested in groups of from 1 to 26 in size. Each rater was given a packet containing his 250 profiles and was told that these were profiles of men in the level immediately beneath their own in the same career ladder. They were to examine each profile and give it a score from one to nine in terms of the judged criterion "worth to the Air Force." The actual procedure consisted of having the rater sort the profiles into thirds (high, medium and low) and then sort each third into thirds again. This final sorting had to conform to a forced distribution as follows:

Stanine Score	Number of Profiles
9	10
8	17
7	30
6	43
5	50
4	43
3	30
2	17
1	10

The use of the forced distribution served to maintain equality of means and variances on the criterion for all raters.

Results

Regression analyses were performed on the judgments of all raters using the profile trait scores as predictors. These regression equations were then clustered using JAN.

Regression Equations

Tables 3 and 4 present the results of the regression analyses. Table 3 presents the zero order correlation between the profile traits and the criterion judgments of the rater. The last column in the table gives the R^2 value for that rater—an expression of the consistency of rater judgment across the 250 profiles. While in general all raters showed a great deal of consistency, there were, of course, substantial differences between raters. The R^2 values ranged from a high of .973 (rater 37) to a low of .569 (rater 14).

The raw score regression equations obtained for each of the raters are given in Table 4 showing the regression weights associated with each of the predictor variables.

Clustering Procedure

The JAN clustering procedure was set up to minimize, at each stage in the clustering process, the change in the composite R^2 (ΔR_c^2) for the entire group where $R_{c_i}^2 - R_{c_j}^2 = \Delta R_c^2 \cdot R_{c_i}^2$ is defined as the composite group predictive efficiency at stage i in the clustering process and $R_{c_j}^2$ as the composite group predictive efficiency at stage j , where $j = i + 1$. Further, the composite predictive efficiency at any stage i ($R_{c_i}^2$) was defined by the objective function

$$R_{c_i}^2 = \frac{SS \text{ reg}_1(n_1) + SS \text{ reg}_2(n_2) + \dots + SS \text{ reg}(k-i)(n_{k-i})}{SS_{\text{total}_1} + SS_{\text{total}_2} + \dots + SS_{\text{total}_k}} \quad (3)$$

where k is the original number of regression equations to be grouped and i is the number of stages or groupings made—thus $k - i$ is the number of groups that exist at that particular stage. Note that the denominator of (3) remains constant throughout the entire cluster-

TABLE 3
Zero-Order Validity Coefficients (r_i 's) Associated with Each of the Profile Traits for all Raters

Rater	Trait																							R
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
1	526	450	476	537	639	607	616	439	701	651	728	680	527	611	433	500	519	684	584	520	478	509	385	880
2	389	209	170	498	279	352	431	337	375	729	344	334	188	303	179	343	269	649	749	605	344	218	866	866
3	513	437	442	514	623	573	576	405	668	693	683	665	457	574	411	517	441	700	582	518	465	435	454	823
4	436	439	329	473	486	408	521	388	518	585	555	474	407	442	372	433	413	569	619	490	443	342	437	648
5	567	426	364	471	555	461	524	508	620	548	620	560	448	512	403	509	581	543	501	506	557	486	342	776
6	561	486	405	562	612	464	570	436	599	597	691	601	509	589	403	571	534	577	581	487	529	419	376	818
7	503	288	261	506	421	372	585	605	501	559	532	483	448	465	283	485	514	544	609	583	562	495	414	748
8	565	173	180	403	321	246	403	480	438	387	444	418	458	353	146	385	748	356	423	376	472	634	232	762
9	469	184	271	443	355	324	536	547	472	394	482	473	439	580	190	608	418	429	453	386	456	364	259	684
10	576	456	440	523	596	542	566	410	679	679	694	632	468	599	458	556	525	661	581	551	511	447	415	860
11	518	445	475	503	583	636	594	347	705	726	672	697	425	592	380	498	452	703	576	515	441	435	411	839
12	541	477	414	466	588	535	543	434	626	604	665	623	487	581	473	475	533	609	489	448	508	409	346	775
13	688	271	183	325	386	268	350	397	484	410	507	411	466	453	244	438	890	334	425	297	502	397	247	890
14	396	418	394	485	528	497	475	302	545	526	512	542	374	448	328	419	339	545	478	461	380	282	353	569
15	520	554	478	508	674	513	555	429	694	623	722	642	475	592	538	513	513	605	500	484	486	429	332	854
16	537	349	386	445	507	499	533	439	647	544	608	620	476	470	358	462	553	578	499	450	493	672	323	798
17	515	368	333	529	519	433	599	442	560	560	540	493	537	328	488	609	564	611	509	549	453	404	775	
18	537	185	215	367	361	297	557	406	507	390	507	505	447	527	205	433	763	395	404	347	469	398	182	757
19	553	359	384	499	509	547	564	411	662	642	639	632	451	578	350	461	573	634	562	509	521	482	428	792
20	568	409	334	507	540	446	627	428	604	561	641	586	502	561	383	511	534	551	515	462	637	405	322	791
21	492	452	406	561	587	602	599	396	675	711	695	645	372	615	420	471	418	730	583	588	453	391	469	841
22	424	309	308	521	453	442	717	328	509	565	544	548	420	494	229	444	411	562	573	484	418	367	378	680
23	490	333	308	571	531	485	830	408	593	584	616	635	534	643	286	550	458	610	601	506	456	387	362	889
24	613	353	333	589	534	443	680	455	551	635	641	564	534	533	316	571	543	584	650	582	564	430	420	845
25	513	406	380	507	546	535	570	403	633	612	649	605	423	524	405	484	489	650	566	536	413	408	401	768

TABLE 3 (Continued)

Rater	Trait																							R ²
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
26	485	561	527	485	722	657	526	321	693	660	744	695	462	591	535	458	423	645	475	423	408	395	341	830
27	573	497	437	529	569	554	514	399	608	662	645	571	393	533	533	457	482	471	632	570	510	443	390	484
28	594	378	398	475	545	490	589	382	680	594	664	647	531	606	359	518	652	571	506	417	479	470	325	820
29	476	355	310	630	500	472	609	415	588	687	553	536	373	465	309	446	398	674	746	647	446	443	582	832
30	501	458	460	531	628	636	612	436	734	691	721	708	504	628	460	493	522	690	562	527	500	486	389	890
31	422	453	358	261	508	487	404	284	619	480	575	552	385	464	390	316	473	479	323	324	428	450	205	590
32	475	325	306	446	435	445	502	445	497	486	538	492	408	478	304	435	512	544	483	502	452	373	293	603
33	601	300	296	421	451	397	496	458	583	508	582	552	494	507	281	460	769	479	457	389	549	571	255	816
34	485	527	483	509	651	634	552	394	734	696	754	694	440	651	498	493	502	671	526	510	463	451	360	877
35	556	397	355	458	515	436	467	554	548	495	623	502	430	499	351	437	679	499	481	473	586	484	313	777
36	527	475	480	560	617	606	621	452	667	711	735	660	447	592	435	518	504	698	614	562	505	450	427	884
37	443	076	106	529	320	265	981	297	410	404	407	516	567	440	040	539	359	473	566	466	403	324	301	973
38	537	484	432	549	611	541	545	414	622	594	665	611	499	546	433	502	486	583	519	473	474	429	333	759
39	544	429	390	473	566	493	536	416	604	608	648	572	467	553	394	490	507	607	551	468	426	425	422	731
40	496	279	285	428	416	425	608	492	561	567	587	577	404	503	266	452	496	535	510	452	427	385	352	671
41	519	294	180	547	366	297	514	670	406	530	414	328	369	341	262	453	499	483	610	544	513	334	471	734
42	549	409	328	466	519	495	490	484	577	566	689	556	446	508	335	487	582	538	530	417	483	429	325	730
43	480	539	498	527	637	594	514	414	643	663	682	624	417	548	474	432	474	629	522	534	484	433	353	797
44	537	501	404	533	677	558	600	459	625	665	726	608	507	568	474	485	521	659	617	515	493	457	393	862
45	526	482	398	528	583	495	564	487	588	591	631	572	469	557	459	512	545	575	551	498	529	429	378	788
46	515	436	493	523	634	634	587	371	699	721	727	679	430	624	460	461	466	722	591	525	449	491	426	870
47	454	153	188	568	382	344	934	341	466	476	471	552	568	496	122	599	335	538	587	502	453	335	335	915
48	595	462	440	520	625	562	581	484	684	615	705	645	506	584	454	560	585	608	566	521	528	509	377	882
49	512	504	491	535	664	651	578	396	692	694	721	688	484	605	481	502	451	686	576	508	467	436	402	863
50	520	410	409	544	572	581	577	389	683	674	669	676	455	589	355	510	485	658	581	475	456	425	425	798

TABLE 4
Raw Score Regression Weights (b_i 's) Associated with Each of the Profile Traits for all Raters

Rater	Trait																							Con- stant
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
1	068	000	136	081	072	046	044	049	163	000	149	000	124	081	000	000	061	076	126	170	000	077	000	-2.627
2	000	000	000	000	000	085	000	105	000	000	000	174	000	000	048	000	000	000	142	000	035	000	736	-1.654
3	000	000	070	000	198	000	000	078	113	133	000	080	061	000	000	159	059	190	082	050	096	000	058	-2.141
4	000	206	039	000	000	000	080	059	167	000	000	000	035	000	057	088	040	069	310	000	114	000	049	-1.537
5	146	071	000	000	147	000	057	165	230	000	000	000	000	051	038	074	111	000	000	136	099	087	064	-2.378
6	000	159	063	130	051	-065	000	052	133	000	163	050	047	092	000	139	106	000	183	000	114	000	050	-2.359
7	000	000	000	000	078	000	048	279	000	000	000	195	000	000	043	092	079	000	170	174	122	100	049	-2.116
8	128	000	000	131	000	000	000	052	000	000	000	000	047	000	000	000	480	000	000	079	000	318	000	-1.168
9	000	000	000	000	000	000	052	360	075	000	000	000	000	329	000	349	000	000	000	077	000	000	000	-1.164
10	120	051	057	000	000	000	000	000	182	160	118	000	000	088	079	170	056	000	059	225	101	043	000	-2.567
11	061	047	074	000	100	087	090	000	189	218	000	074	000	000	-043	140	042	059	072	095	091	000	000	-2.011
12	156	114	048	000	000	000	132	083	052	000	016	113	058	062	156	000	127	172	000	000	126	000	058	-2.384
13	300	081	000	000	000	000	000	000	000	000	000	000	000	088	000	000	687	048	000	000	000	000	000	-1.007
14	000	074	097	136	177	000	000	000	228	000	-095	102	000	000	000	103	000	000	152	114	067	000	044	-1.020
15	049	144	000	050	071	064	093	070	199	085	030	000	000	070	136	143	084	000	000	119	073	050	000	-2.649
16	148	056	043	000	105	000	064	000	192	000	000	000	044	000	000	039	072	050	000	168	060	361	048	-2.260
17	000	058	053	081	000	000	149	000	129	000	155	000	049	050	000	000	231	000	207	048	107	043	044	-2.014
18	056	000	000	000	000	000	291	070	000	000	000	072	-046	179	000	000	596	000	-039	000	000	000	000	-0.850
19	117	000	054	000	032	144	062	000	168	059	000	000	057	120	000	000	166	000	102	112	120	066	076	-2.283
20	156	103	000	000	000	000	271	000	135	118	147	000	000	000	059	000	080	000	000	000	298	000	000	-1.871
21	064	076	000	095	000	000	050	045	169	000	202	000	000	097	051	116	000	220	000	164	000	000	086	-2.193
22	000	045	081	119	000	000	000	045	000	169	110	000	000	000	000	000	118	000	079	000	000	047	000	-1.137
23	000	069	000	066	051	037	545	098	066	000	088	000	000	188	000	038	074	000	000	000	000	000	120	-2.182
24	164	000	067	065	054	000	217	000	000	163	193	000	043	000	000	071	086	-042	055	175	105	047	000	-2.333
25	040	051	000	048	112	049	081	000	087	000	000	000	000	000	067	119	148	219	000	127	086	162	052	-2.225

TABLE 4 (Continued)

Trait

Rater	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Con- stant
26	109	078	089	104	176	057	128	000	141	161	115	061	041	000	051	000	000	000	000	000	041	000	042	-2.016
27	208	136	098	111	000	055	000	053	112	072	087	000	000	061	067	104	039	000	000	094	000	069	178	-2.742
28	124	000	097	053	113	000	178	000	218	098	000	000	039	073	000	053	288	000	000	000	000	038	038	-2.075
29	000	000	000	178	138	000	000	000	285	000	000	000	000	000	000	046	000	000	328	176	000	124	104	-1.900
30	000	000	041	047	134	081	074	086	231	089	000	078	074	081	046	038	054	000	109	073	098	038	043	-2.604
31	120	282	000	-077	000	000	076	000	277	000	000	000	000	000	000	000	100	095	000	000	153	142	000	-0.866
32	057	096	000	000	000	090	000	099	-042	-087	000	103	039	056	000	141	231	246	000	214	038	000	000	-1.368
33	142	000	000	000	082	000	079	053	000	124	000	159	000	000	000	000	448	000	000	000	111	135	000	-1.685
34	000	107	000	054	042	000	000	000	221	192	178	000	000	121	051	127	049	000	000	164	091	064	000	-2.347
35	083	062	051	000	175	000	000	174	000	-046	000	078	000	040	000	000	329	118	000	070	179	073	045	-2.136
36	044	056	115	054	000	000	092	059	097	136	247	066	000	000	000	000	106	000	000	174	085	000	000	-2.539
37	000	000	000	000	000	000	000	000	000	000	000	000	037	000	000	000	106	000	000	000	000	000	000	-0.485
38	141	108	085	170	065	000	000	086	143	087	097	000	097	063	000	072	041	000	000	121	000	069	000	-2.271
39	127	071	084	000	069	000	000	092	146	000	101	000	077	097	000	072	074	064	087	108	000	055	126	-2.250
40	103	000	000	000	000	000	295	222	056	190	176	114	-042	000	000	000	145	-052	000	000	000	000	046	-1.280
41	112	128	000	130	000	000	109	443	093	000	-078	000	000	000	059	065	103	000	074	000	000	000	170	-1.998
42	101	073	000	000	000	066	000	174	000	061	300	128	000	000	-051	065	175	000	174	000	045	000	000	-1.585
43	110	119	112	109	088	000	044	000	143	132	153	000	000	000	000	000	044	000	000	201	118	079	000	-2.298
44	083	069	000	000	183	000	112	110	105	101	149	000	070	000	052	000	063	059	173	051	040	050	000	-2.355
45	000	186	043	078	000	000	146	127	125	072	001	000	000	066	107	130	150	000	000	079	124	052	060	-2.719
46	098	000	105	073	079	046	144	000	092	164	163	000	000	091	041	000	040	047	000	108	044	099	067	-2.518
47	000	045	000	045	000	000	753	000	000	000	000	000	038	000	000	139	000	113	000	000	076	000	000	-1.002
48	162	045	063	000	095	000	072	103	165	000	084	046	000	091	084	099	105	000	000	181	038	083	087	-3.024
49	049	047	079	091	097	128	000	044	111	179	116	000	103	064	000	082	000	000	075	113	067	046	000	-2.487
50	000	000	000	104	144	072	000	046	190	128	000	134	000	059	000	110	103	000	158	000	087	000	040	-1.927

ing process; only the numerator changes. Thus JAN actually minimizes the drop in SS regression that occurs when a single equation is used to express the variation in two groups, each of which has its own equation. For example, given two groups A and B at stage i , each group has its own equation and SS regression. If at stage j these groups are combined and a single equation is used to describe the data points the resulting SS reg will always be equal to or less than $SS \text{ reg}_A + SS \text{ reg}_B$. Thus, the expression $SS \text{ reg}_A + SS \text{ reg}_B - SS \text{ reg}_{\text{combined data}}$ is a measure of the loss in predictive efficiency occurring by combining these two groups. JAN computed which grouping resulted in the smallest change in the numerator of (3) at each stage.

The R_c^2 values as the number of clusters was systematically reduced from 50 separate equations to a single overall equation in steps of one are shown in Table 5. The process can be seen even more graphically in Figure 2. Table 5 also shows the results of applying the F test given in equation Z upon each successive ΔR^2 value throughout the clustering process. Since the degrees of freedom for all F 's were, for all practical purposes, 23 and infinity rejection values of 1.55 ($p < .05$) and 1.84 ($p < .01$) were employed. Using the more conservative probability one finds the ΔR_c^2 values becoming significant at the 25th stage, indicating there are 25 clusters or rater policies which can be considered significantly different from each other.

Greater insight into the clustering of the raters can be obtained from examination of Figure 3, which presents the grouping of raters in sequential form. The numbers in each box indicate the number of groups remaining after that combination has been made. At the far right, the final grouping is designated by a 1, indicating that with this grouping all raters were now combined into a single composite or cluster. Following the lines extending from this junction, one can trace the grouping process. Thus, the last grouping joined together a small group of 5 raters (Numbers 8, 13, 18, 33, and 35) with the larger cluster consisting of the remaining 45 supervisors. This "branching out" process can thus be traced throughout the entire "tree" down to the last grouping (or the first, depending upon your point of view) which is indicated at the far left with number 49. At that point there were 49 clusters—48 single rater policies and one policy representing a pair of raters.

TABLE 5
*F Tests between Composite R^2 Difference Values
 at Each Stage in the Grouping Process*

No. of groups remaining	Stage	R_o^2	$R_{o_i}^2 - R_{o_f}^2$	df	F
50	0	8041	—	—	—
49	1	8039	0002	23/11,350	.5055
48	2	8037	0002	23/11,373	.5044
47	3	8035	0002	23/11,396	.5054
46	4	8032	0003	23/11,419	.7596
45	5	8029	0003	23/11,442	.7562
44	6	8027	0003	23/11,465	.5084
43	7	8024	0003	23/11,488	.7592
42	8	8021	0003	23/11,511	.7607
41	9	8017	0004	23/11,534	1.0130
40	10	8013	0004	23/11,557	1.0150
39	11	8010	0004	23/11,580	.7602
38	12	8006	0004	23/11,603	1.0140
37	13	8002	0004	23/11,626	1.0160
36	14	7997	0005	23/11,649	1.2662
35	15	7992	0005	23/11,672	1.2687
34	16	7988	0005	23/11,695	1.0119
33	17	7983	0005	23/11,718	1.2635
32	18	7977	0005	23/11,741	1.5161
31	19	7972	0006	23/11,764	1.2634
30	20	7966	0006	23/11,787	1.5169
29	21	7960	0006	23/11,810	1.5148
28	22	7953	0007	23/11,833	1.7647
27	23	7946	0007	23/11,856	1.7629
26	24	7939	0007	23/11,879	1.7612
25	25	7931	0008	23/11,902	2.0078
24	26	7923	0008	23/11,925	2.0065
23	27	7914	0008	23/11,948	2.2493
22	28	7905	0009	23/11,971	2.2433
21	29	7896	0009	23/11,994	2.2424
20	30	7887	0009	23/12,017	2.2362
19	31	7877	0010	23/12,040	2.4761
18	32	7867	0010	23/12,063	2.4703
17	33	7855	0012	23/12,086	2.9532
16	34	7843	0013	23/12,109	2.9430
15	35	7830	0013	23/12,132	3.1807
14	36	7816	0014	23/12,155	3.4087
13	37	7802	0014	23/12,178	3.3940
12	38	7785	0016	23/12,201	4.1006
11	39	7766	0020	23/12,224	4.5601
10	40	7742	0023	23/12,247	5.2449
9	41	7718	0024	23/12,270	5.6709
8	42	7692	0027	23/12,293	6.0877
7	43	7664	0027	23/12,316	6.4954
6	44	7635	0029	23/12,339	6.6577
5	45	7585	0050	23/12,362	11.3623
4	46	7517	0068	23/12,385	15.1636
3	47	7426	0091	23/12,408	19.7719
2	48	7249	0177	23/12,431	37.1634
1	49	7067	0182	23/12,454	35.8243

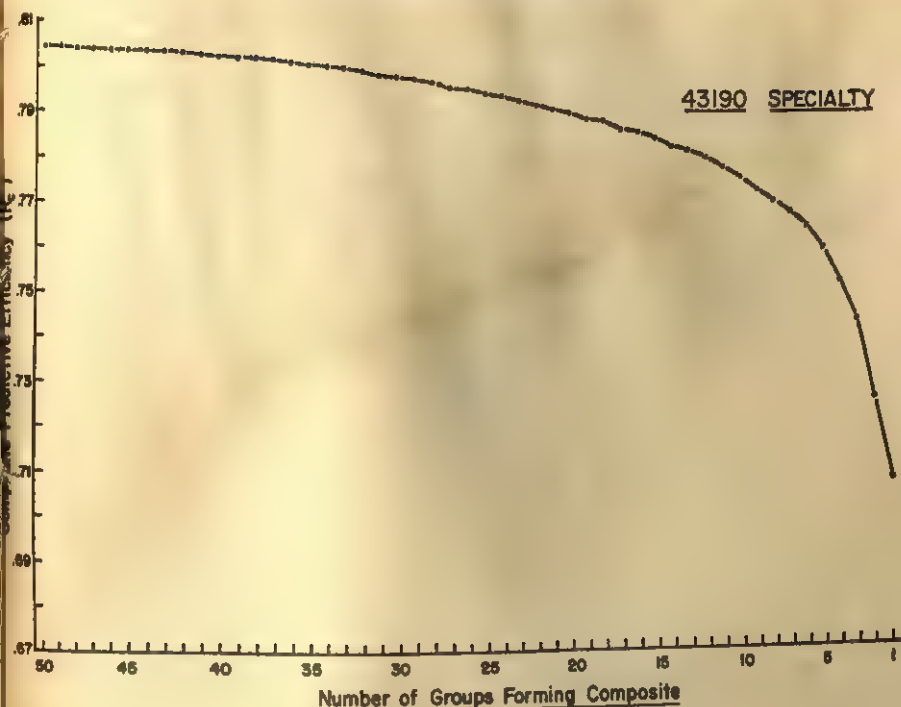


Figure 2. Composite multiple correlation (R_c^2) values as the number of groups of raters is systematically reduced from fifty to one by JAN.

Discussion

The use of simulated job incumbent profiles generated using the basic factor analysis model appeared to be a successful procedure for obtaining judgments from raters concerning job attributes—a particularly gratifying result when one considers the underlying structure used to generate profiles was one obtained by logical but arbitrary decisions of experts rather than one derived empirically. That the profiles were indeed perceived as being meaningful and representing real people was indicated in several ways. During the initial generation of the profiles supervisors in the specialty involved were asked to examine them to see if there were any which they felt were “unreasonable” in that the scores were not realistic. Only four such profiles were identified among several thousand generated! Also, not one comment was received from any of the 50 raters which might lead the authors to believe that a profile was felt to be un-

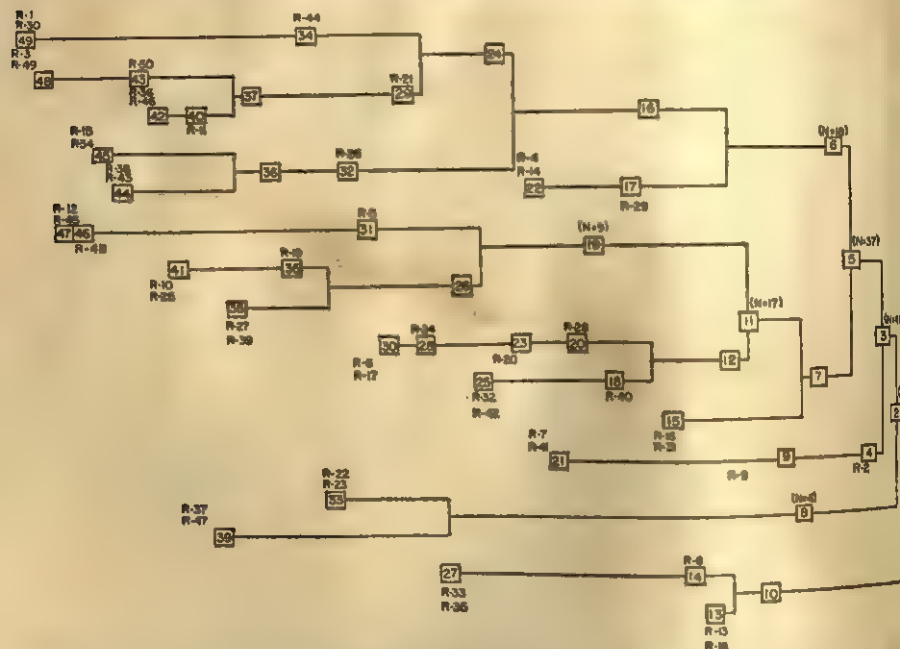


Figure 3. Tree diagram showing a pictorial view of the sequential clustering of fifty raters from fifty individual groups down to a single composite group.

reasonable in terms of its scores or that it did not represent a real person.

The advantage of being able to use simulated data with JAN removes one of the major problems associated with the method—that of having to obtain large numbers of stimuli which have been quantified on each of the dimension assumed to have some potential relevance to the judgment dimension of raters. The potential savings in time and money made possible by artificially obtained stimuli would seem to have really tremendous implications—particularly where the stimuli are people and where one is interested in deciding which traits among a large number *need* to be measured. One unanswered question, of course, is the degree to which variations in the arbitrary R matrix can lead to variations in the experimental outcomes—a question currently under investigation. Certainly it is evident that the stimuli used in this study resulted in highly consistent intra-judge performance as indicated by the generally large R^2 values for all raters. To be able to account for this much variance in a judge's categorization of people in terms of an overall criterion

of "worth to the Air Force" would certainly seem to be a rewarding result. Indeed, the R^2 values obtained for the raters compared quite favorably with those reported by Madden (1963).

It is interesting to find that those traits which appeared to predict the global rating best, such as *willingness to work*, *maintaining discipline*, etc., were generally supervisory in nature, which some of the traits felt to be of much less importance were those involving job knowledge and skill, e.g., *knows theory of aircraft systems*. By and large, those traits which might be thought of as relating to leadership were those in which the raters seemed to place the greatest reliance—a not unsurprising outcome since the specialty under study was supervisory in nature by Air Force definition.

The F test, when used as suggested by Bottenberg and Christal (1961) seemed to work—at least it indicated that there were 25 statistically different groupings of raters, or that there were 25 statistically different rater policies represented among the 50 raters. The question of how meaningful these differences were (i.e., the practical significance question) is another matter. However, the fact that, even when a single regression equation was used to explain the judgments of all 50 raters, an R_c^2 of .707 was obtained certainly casts doubt on the practical importance of the significance findings. After all, only a 10 percent loss in predictive efficiency was observed throughout the entire clustering process (80 percent-70 percent), thus the rater equations were basically quite homogeneous.

REFERENCES

- Bottenberg, R. A. and Christal, R. E. *An Iterative Technique for Clustering Criteria which Retains Optimum Predictive Efficiency*. Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Division, March 1961. (WAAD-TN-61-30, ASTIA Document AD-261 615).
- Christal, R. E. *JAN: A Technique for Analyzing Individual and Group Judgment*. Lackland Air Force Base, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, February, 1963. (PRL-TDR-63-3, ASTIA Document AD-403 813).
- Cronbach, L. J. and Gleser, G. C. "Assessing Similarity between Profiles." *Psychological Bulletin*, L (1953), 456-473.
- Haggard, E. A., Chapman, Jean P., Isaacs, K. S., and Dickman, K. W. "Intraclass Correlation versus Factor Analytic Techniques for Determining Groups of Profiles." *Psychological Bulletin*, LVI (1959), 48-57.
- Madden, J. M. *An Application to Job Evaluation of a Policy-Capturing Model for Analyzing Individual and Group Judgment*.

- ments. Lackland Air Force Base, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, May 1963. (PRL-TDR-63-15)
- McQuitty, L. L. "Hierarchical Linkage Analysis for the Isolation of Types." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XX (1960), 55-67.
- McQuitty, L. L. "Isolating Predictor Patterns Associated with Major Criterion Patterns." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVII (1957), 3-42.
- Naylor, J. C. and Wherry, R. J. *Feasibility of Distinguishing Supervisors' Policies in Evaluation of Subordinates Using Ratings of Simulated Job Incumbents*. USAF PRL Technical Documents Report, No. 64-25, October, 1964.
- Nunnally, J. "The Analysis of Profile Data." *Psychological Bulletin*, LIX (1962), 311-319.
- Sawrey, W. L., Keller, L., and Conger, J. J. "An Objective Method of Grouping Profiles by Distance Function and its Relationship to Factor Analysis." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XX (1960), 651-674.
- Ward, J. H. *Hierarchical Grouping to Maximize Payoff*. Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Division, March, 1961. (WADD-TN-61-29, ASTIA Document AD-261 750)
- Wherry, R. J., Sr., Naylor, J. C., Wherry, R. J., Jr. and Fallis, R. F. "Generating Multiple Samples of Multivariate Data with Arbitrary Population Parameters." *Psychometrika*, (In Press).

NUMBER OF SCALE POINTS AND THE RELIABILITY OF SCALES

S. S. KOMORITA AND WILLIAM K. GRAHAM

Wayne State University

Among the standard scales and inventories in current use, there seems to be relatively little agreement regarding the number of scale points or categories which are used to obtain responses to items. For example, in the California Psychological Inventory (Gough, 1957), a true-false, two-point scale is used; in a Likert-type attitude scale (Likert, 1932), a five-point scale is employed; and in the California F-scale (Adorno, et al, 1950) or the Semantic Differential (Osgood, Suci, and Tannenbaum, 1957), a seven-point scale is employed. Clearly, scales which employ a two-point scale such as agree-disagree, true-false, etc., are much shorter and more convenient to administer and score. If so, why do so many investigators recommend and use the seven-point scale?

One obvious criterion for the choice of the number of scale categories is the ability of subjects to discriminate between the categories. While a scale with few categories may not allow the subject to make full use of his capacity to discriminate, a scale with a large number of categories may be beyond the subject's capacity to discriminate, thus, increasing errors of measurement. Thus, probably the main reason for using a large number of categories is to increase the reliability of the scale. As applied to rating scales, in one of the earliest studies of this problem, Symonds (1924) made a theoretical analysis using Kelley's correction for coarse grouping on the correlation coefficient, and concluded that the optimal number of categories to maximize scale reliability is seven, and that the increase in reliability when more categories are used is negligible. However,

more recent empirical studies on this problem have not supported Symonds' conclusions.

Champney and Marshall (1939) compared the correlation between two forms of a graphic rating scale and measured responses to each form by two methods: using a coarse centimeter scale and a finer millimeter scale. The correlation between the two forms for the millimeter scale was significantly higher than for the centimeter scale, .77 versus .67. They concluded that the optimal number of scale categories is a function of the conditions of measurement and cannot be determined by the method proposed by Symonds. Guilford (1954, p. 291) is in general agreement with this point of view and claims that, "the number 7 recommended by Symonds is usually lower than optimal and it may pay in some favorable situations to use up to 25 scale divisions." On the other hand, two studies by Bendig (1953, 1954) suggest that fewer than seven categories may be justified under certain conditions. In the first of these studies, he found that reliability remained relatively constant for self-rating scales with 3, 5, 7, and 9 categories but significantly decreased for 11 categories. In the second study, Bendig confirmed the results of his first study and found no significant differences in the reliability of rating scales with three to nine categories, but found that a two-point rating scale was significantly lower in reliability than those with three to nine categories. Thus, his results suggest that although there is a definite advantage in using more than a two-point scale, it may be possible under certain conditions to use fewer than seven categories without sacrificing reliability.

In summary, the over-all results of studies concerning the effects of number of scale points on the reliability of rating scales indicate that in some situations *more* than seven categories are optimal, while in other situations, *fewer* than seven categories may be justified. Thus, the effect of number of scale points on reliability seems to vary with the stimulus situation. The major implication, of course, is that the problem is much more complex than simply applying a correction for coarse grouping.

It should be emphasized, at this point, that the studies which have been reviewed dealt primarily with the relationship between number of scale points and the reliability of a *single rating scale*. The purpose of the present study, however, was to determine the effects of number of scale points on the reliability of inventories and test-scales which consist of the *sum of a set of rating scales* (e.g., Likert's

method of summated ratings). In one of the few studies dealing specifically with this problem, Bendig (1954) obtained ratings of food preferences for 20 different foods and pooled the 20 ratings for each subject. Five rating scales with 2, 3, 5, 7, and 9 scale categories were administered to five groups of subjects, each group receiving one of the scales. Using Hoyt's analysis of variance technique as a measure of reliability, the reliabilities of the five scales ranged from .60 to .70, and he concluded, therefore, that test reliability is independent of the number of scale categories.

Similar results were obtained in a study by Komorita (1963). Two forms of a Likert-type attitude scale consisting of 14 items, each with a six-point scale, were administered to a sample of 286 subjects. Responses to the two forms were scored using a two-point scale as well as using the six-point scale. For the six-point scale, the correlation between the two forms was .93, while the correlation for the two-point scale was .91, thus, confirming Bendig's results. In a supplementary study, the same comparison was made for two random samples of three items from the original scales. For the six-point scale, the correlation between the two sets of three items was .83, while the correlation for the two-point scale was .71. Since the difference in reliabilities between the six-point and two-point scales was considerably larger for three items than for the 14 items, it was suggested that if a scale consists of a very small number of items, somewhat better reliability might be obtained if a six or seven-point scale is used instead.

The over-all results of the studies by Bendig and by Komorita indicate that test reliability is independent of number of scale points. If this is a valid generalization, the major implication is that, because of simplicity and convenience in administration and scoring, all inventories and scales ought to use a dichotomous, two-point scoring scheme. Similar conclusions have been made by Peabody (1962) who attempted to analyze scores on bi-polar scales into two components: direction of response and intensity or extremeness of response. His results indicated that composite scores consisting of the sum of scores on bi-polar, six-point scales mainly reflect direction of response, and are minimally influenced by extremeness of response. He concluded, therefore, that there is justification for scoring bi-polar items dichotomously according to direction of response.

There are some reasons to believe, however, on both empirical and

theoretical grounds, that a number of variables may affect the generality of this principle. It is plausible, for example, that the number of items in the scale or the homogeneity of the items in the scale might affect the relationship between test reliability and number of scale points; moreover, there might be an interaction between these variables in their effects on test reliability. It is desirable, therefore, to specify in more detail the conditions under which test reliability is independent of number of scale points. Accordingly, the primary purpose of this study was to determine the effects of two variables, number and homogeneity of items, on the relationship between test reliability and number of scale points.

Method

Scales

To determine the effects of homogeneity of item content, two scales were selected known to vary in homogeneity of content: (a) a relatively homogeneous scale consisting of 24 bi-polar, Semantic Differential adjectives (Osgood, et al, 1957) selected for their high loadings on the evaluative factor (*Ss* were asked to rate Gov. George Romney of Michigan), and (b) a random sample of 24 items from the sociability subscale of the California Psychological Inventory (Gough, 1957) whose internal consistency reliability is reported to be approximately .70.

The sociability scale, hereafter referred to as the CPI scale, and the 24 bi-polar adjectives, hereafter referred to as the SD scale, were each presented in two forms. In Form A, the items were presented with a two-category scale, while in Form B, the items were presented with a six-category scale. Except for differences in number of scale categories, the two forms were identical.

Subjects

The *Ss* were 260 students enrolled in undergraduate psychology courses. Forms A and B of the CPI were administered to 67 and 56 *Ss*, respectively, and Forms A and B of the SD were administered to 67 and 70 *Ss*, respectively.

Procedure

The plan of the study was to compare the difference in reliability between the two-point and six-point scales for the SD and CPI

scales. For this purpose, Cronbach's coefficient alpha (1951) was determined for each form. In order to minimize errors of coarse grouping for the two-category interitem correlations, each form was randomly partitioned into four subsets of six items and coefficient alpha was computed with k equal to four.

To determine the effects of number of scale points on reliability as a function of number of items, the reliabilities of the scales for 3, 6, 12, and 36 items were estimated by applying the Spearman-Brown formula to the previously-determined alpha coefficients. As an empirical check on the assumptions of the Spearman-Brown estimates, for each form the mean intercorrelations between the four random subsets of six items was compared with the theoretical Spearman-Brown estimates. In addition, each form was also partitioned into eight random subsets of three items and two random subsets of 12 items (split-half), and the empirical values were compared with the Spearman-Brown estimates.

Results

Table 1 shows the alpha coefficients for each of the forms. It can be seen that the difference between the two-point and six-point SD scales is negligible, while the difference between the two-point and six-point CPI scales is moderately large. This differential effect of number of scale points as a function of the homogeneity of the scale is further demonstrated in Figure 1. This figure shows the

TABLE 1
Alpha Coefficients for Two-Point and Six-Point SD and CPI Scales

	SD		CPI	
	2-pt. 67	6-pt. 70	2-pt. 67	6-pt. 56
n				
a	.920	.916	.620	.740

theoretical Spearman-Brown estimates of reliability as a function of number of items as well as the empirical values for 3, 6, and 12 item subscales. For three and six items, the empirical values represent the mean intercorrelations between eight and four random subscales, respectively, while for 12 items, the empirical values represent a single split-half correlation. It can be seen that an excellent fit between theoretical and empirical values was obtained, thus,

justifying the use of the Spearman-Brown formula. Figure 1 also shows that the differences between the two-point and six-point SD scales are negligible across number of items (in fact, the reliabilities for the two-point form are slightly higher than for the six-point form), while the reliabilities of the six-point CPI subscales are *consistently* higher than for the two-point CPI subscales.

Finally, it should be noted that differences between the two-point

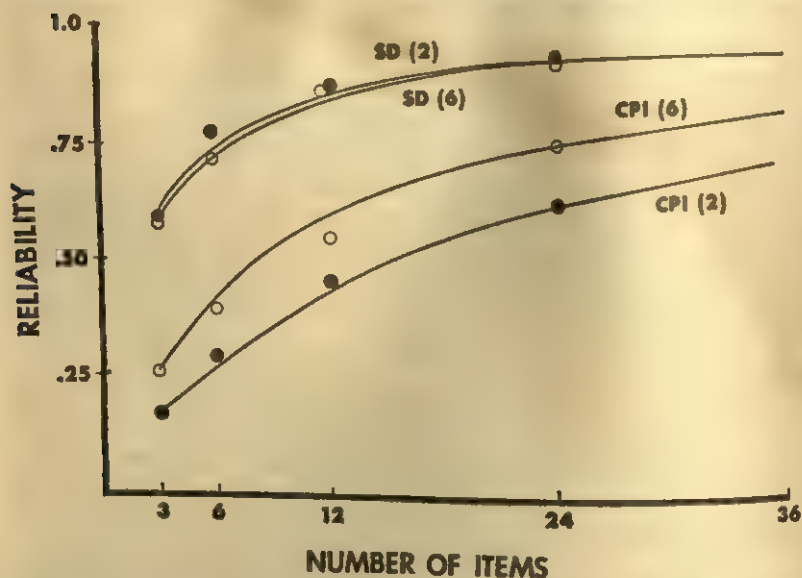


Figure 1. Reliability as a function of number of items (curves represent theoretical Spearman-Brown estimates while points represent empirical values).

and six-point forms, as function of number of items, are relatively constant for *both* the SD and the CPI scales. Although this result would suggest that number of items in the scale has no effect on the relationship between number of scale points and reliability, theoretically at least, the reliability of the scales should converge to 1.0 as the number of items becomes indefinitely large. Thus, it is quite plausible that if the number of items in the CPI scale were to be made indefinitely large, the *absolute difference* between the two-point and six-point scales would be much smaller. For example, if the CPI scales were to be increased to 240 items, the Spearman-Brown estimates for the two-point and six-point scales would be .942 and .965, respectively, and this difference is considerably smaller than the difference of .620 and .740 for the 24 item scales.

Hence, it is reasonable to assume that if a much larger number of items had been used, the results would have indicated that the effects of number of scale points on reliability would become negligible as the number of items in the scale became indefinitely large.

Discussion and Conclusions

For the SD scale, the results of this study are consistent with the results found by Bendig (1954) and by Komorita (1963), and indicate that with a relatively homogeneous set of items, the reliability of a scale is independent of the number of item scale points. If the items are relatively heterogeneous, however, the results suggest that the reliability of the scale can be increased not only by increasing the number of items but also by increasing the number of item scale points. Just how heterogeneous the scale must be before one can expect a reasonable increase in reliability is a matter for further research. The results of this study suggest that if the reliability of a scale with approximately 25 items and a two-point scale is approximately .60, if the number of scale points is increased to six, one can expect an increase in reliability of about .15.

The major implication of this result is that if there are reasons to believe that the items in a proposed scale or inventory can be expected to be homogeneous, either by item analysis selection of items as in Likert's technique of scaling or by a factor analysis of a set of items as in the Semantic Differential, then a two-category response format such as true-false, agree-disagree, etc., will yield as high a reliability coefficient as a multi-category system. In terms of ease of administration and scoring, therefore, a two-point scheme seems to be preferable to a multi-category scheme.

On the other hand, even with a homogeneous scale, the use of a multi-category system, under certain conditions, may be justified. In using Guttman's scale analysis (1950), for example, an investigator may wish to obtain a separate intensity score as well as a content score using what Suchman has described as the "foldover technique" (1950). Moreover, with a two-point format and an extremely small number of items, the variability of the measure would be quite limited, and it may be desirable to use a multi-category response scheme to increase variability.

It should be strongly emphasized, however, that the above discussion certainly is not meant to imply that the decision to use a small

or large number of scale categories should be based solely on the criterion of increasing reliability. The present study as well as most previous studies on this problem has emphasized reliability as the major criterion in the choice of number of scale categories. The ultimate criterion, of course, is the effect on the validity of the scale. In this connection, a reasonable question one could ask is, "with a relatively heterogeneous scale, (and not with a homogeneous scale), why should an increase in the number of scale points increase the reliability of the scale?"

One possible explanation is that an increase in number of scale points increases the precision of the measuring instrument comparable to measuring height in inches rather than in feet or yards. This explanation, however, does not account for the differential effects for homogeneous scales. A more plausible explanation, therefore, is that some type of response set such as an "extreme response set," (Cronbach, 1946; 1950) may be operating to increase the reliability of heterogeneous scales. If the reliability of the response set component is greater than the reliability of the content component of the scale, the reliability of the scale will be increased by increasing the number of scale points. On the other hand, if the reliability of the response set component is less than or equal to the reliability of the content component, as in a homogeneous scale, the reliability of the scale may not be affected or even decreased. Thus, it is reasonable to assume that increasing the number of scale points permits an extreme response set to be evoked, and the use of a two-point response scale eliminates or minimizes this set.

If this interpretation has any validity, a major implication is that the increase in reliability produced by an increase in number of scale points is a spurious one and may or may not increase the validity of the scale. If the response set component correlates with the criterion, the validity of the scale should be increased by the increase in reliability. However, if the response set component does not correlate with the criterion, the validity of the scale should not be affected despite the increase in reliability. Thus, we have an anomolus relation between reliability and validity as in the attenuation paradox (Loevinger, 1954) where an increase in the reliability of a scale may not have any effect on the validity of the scale. Studies are currently in progress to determine the validity of this interpretation.

REFERENCES

- Adorno, T. W., Frenkel-Brunswik, Else, Levinson, D. J. and Sanford, R. N. *The Authoritarian Personality*. New York: Harper, 1950.
- Bendig, A. W. "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on the Scale." *Journal of Applied Psychology*, XXXVII (1953), 38-41.
- Bendig, A. W. "Reliability and the Number of Rating Scale Categories." *Journal of Applied Psychology*. XXXVIII (1954), 38-40.
- Champney, H. and Marshall, Helen. "Rater's Minimal Discrimination as a Criterion for Determining the Optimal Refinement of a Rating Scale." *Journal of Applied Psychology*, XXIII (1939), 323-331.
- Cronbach, L. J. "Response Sets and Test Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 475-494.
- Cronbach, L. J. "Further Evidence on Response Sets and Test Design." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, X (1950) 3-31.
- Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, XVI (1951), 297-334.
- Gough, Harrison G. *Manual, California Psychological Inventory*. Palo Alto: Consulting Psychologists Press, 1957.
- Guilford, J. P. *Psychometric Methods*. (2nd Edition). New York: McGraw-Hill Book Co., Inc. 1954.
- Guttman, L. "The Basis of Scalogram Analysis." In S. A. Stouffer, et al. (Editors) *Measurement and Prediction*. Princeton, New Jersey: Princeton University Press, 1950.
- Komorita, S. S. "Attitude Content, Intensity, and the Neutral Point on a Likert Scale." *Journal of Social Psychology*, LXI (1963), 327-334.
- Likert, R. "A Technique for the Measurement of Attitudes." *Archives of Psychology*, 1932, No. 140.
- Loevinger, Jane. "The Attenuation Paradox in Test Theory." *Psychological Bulletin*, LI (1954), 493-504.
- Osgood, C. E., Suci, C. J., and Tannenbaum, P. H. *The Measurement of Meaning*. Urbana: University of Illinois Press, 1957.
- Peabody, Dean. "Two Components in Bipolar Scales: Direction and Extremeness." *Psychological Review*, LXIX (1962), 65-73.
- Suchman, E. A. "The Intensity Component in Attitude and Opinion Research." In S. A. Stouffer, et al (Editors) *Measurement and Prediction*. Princeton, New Jersey: Princeton University Press, 1950.
- Symonds, P. M. "On the Loss of Reliability in Ratings Due to Coarseness of the Scale." *Journal of Experimental Psychology*, VII (1924), 456-461.



DEVELOPMENT AND CLASSIFICATION OF MODELS FOR MULTIVARIATE ANALYSIS

MARY C. REGAN

University of California, Davis

WHAT is properly called "factor analysis"? One would not expect to find confusion on this point. Yet there appears to be considerable lack of understanding or clarity, for some results are labeled "factor analysis" that in truth are nothing of the sort.

Individuals such as Guttman, Harris, Kaiser, and Rao have continually attacked and systematically resolved elusive problems plaguing the "factor analyst." Their contributions provide an increased understanding of and precision in this particular narrow—but important—section of multivariate methodology.

Regan (1963) classified the various types of analysis developed by these individuals into the following: component type analysis, including complete, incomplete, image and image approximation, and factor analysis. In addition, she applied these various analyses to a particular set of data in order to make explicit the special advantages of each.

Therefore, this paper is divided into three sections. The first section deals with the development of multivariate modes of analysis. The second section classifies these models into component-type and factor-type groupings. The third section presents data illustrating applications of these multivariate models.

Development

Spearman's Two-Factor Model

The important thing in handling multivariate data in terms of underlying factors is the communality concept. Charles Spearman

(1904) introduced this concept when he developed a two-factor model involving a single common factor and a number of specific factors. He asserted that any correlated data could be explained by a single "g" (or general) factor if the tetrad differences of the correlation matrix were zero.

Thurstone's Multiple-Factor Theory

Thurstone (1935), among others, generalized the two-factor Spearman model by postulating that the variables should be analyzed into as many common factors as the rank of the reduced correlation matrix. He saw that a zero tetrad difference (accompanied by the vanishing of higher-order determinants) is the condition for a single underlying factor. Thurstone included Spearman's model as a special case of his own model by considering the reduced correlation matrix as having a rank of one.

The two-factor model initiated by Spearman and generalized into a multiple-factor model by Thurstone was designed to account for a particular matrix by reproducing the correlations, thereby "explaining" the relationships in terms of hypothetical variables called factors.

Factor Analysis

Factor analysis in the Spearman-Thurstone sense first postulates for a set of data a limited number of common factors. This is done by factoring the variance-covariance matrix of common portions through a projective transformation of the original data out of the variable space into the common-factor space. The intercorrelations of the data are then represented by a reduced correlation matrix with a rank equal to the smallest number of common factors that will account for the intercorrelations.

The common-factor space, whose dimensions (m) are defined by the reduced correlation matrix, is distinct from and not embedded in the variable space. Transferral of the vectors from the variable-factor space to the common-factor space reduces their lengths to less than unity by representing only the roots of the communalities.

Hotelling's Principal-Component Model

Hotelling (1933) took a different approach to multivariate analysis of statistical variables. His method of principal components

constructs linear composites of the observed variables by transforming the original set of variables into an uncorrelated derived set that is linear. This unique set of independent factors is equal to the number of variables, completely accounting for the variances and covariances in the given data. The variance is distributed and a portion of the total variance is accounted for by each principal component.

The number of variables can be reduced without affecting the results materially by extracting the factors that carry most of the variance. The factors containing only a small portion of the variance are discarded. When this type of model is reduced in this fashion it is considered to be an incomplete component analysis.

Rao's Canonical Factor Analysis

Because no satisfactory basis existed for determining the limited number of common factors which would reproduce the off-diagonals of a correlation or covariance matrix in factor analysis, Rao (1955) developed the canonical model for factor analysis by deriving the uncorrelated common factors which have the maximum canonical correlations with the observed variables. This model is related to Lawley's (1940) maximum-likelihood solution. Application of the maximum-likelihood solution makes it possible to test statistically whether factors that have been "fitted" are indeed sufficient to account for the data.

The name "canonical factor analysis" was attached to the Rao procedure because it was derived from canonical correlation theory. Rao used a standard canonical analysis to arrive at an equation in the form

$$|R - bU^2| = 0.$$

The solution of the determinantal equation involves R , an inter-correlation matrix with unities in the diagonal, and U^2 , an unknown matrix of unique variances. The unique variances are estimated and then the roots, b , and their characteristic vectors are solved for. In Rao's model, initial estimations of U^2 are modified by an iteration process under the restriction of a given number of common factors.

Harris (1962b, 1963) demonstrated that Rao's procedure provides a means of translating the communality problem into a problem of

rescaling the standardized data by a diagonal matrix, U^{-1} , so as to achieve roots, b_i , of

$$|U^{-1}RU^{-1} - bI| = 0$$

with no b_i less than unity. The advantage is that the canonical correlations are real. Canonical factor analysis is independent of the original, and possibly arbitrary, scale of variables.

Using the Rao procedure for canonical factor analysis makes possible a statistical estimation of the minimum number of factors required. Because Rao's method is independent of any arbitrary numerical test scales or variables, a unique solution can be obtained with either a correlation or covariance matrix. This is also true of Lawley's method.

Guttman's Image Theory

Guttman's (1953) theory postulated that any arbitrary sample of people can be considered a total population and that the test applied can, in reverse, be considered only a sample of a possibly infinite test. He considers variables as an infinite domain or "universe of content." The particular group of variables in any study are then considered a sample arbitrarily selected from this universe. Only linear least-square regressions are then considered, based on an infinitely large population of respondents.

The least-square predicted value of a variable is obtained from the multiple regression of that variable on all the other variables in the data to be analyzed. This predicted value is considered by Guttman to be only a partial image. The total image of a variable is defined as the least-squares predicted value of that variable as determined by the remaining $n - 1$ variables in the *total* universe of content, which constitutes the limit of the partial images as n tends toward infinity.

Guttman's theory holds that any measurement to be made can be partitioned into two independent parts: (1) the image of the variable, and (2) the anti-image of the variable. The image is the predicted value and is dependent on the remaining $n - 1$ variables. The anti-image is the value that is *not* predictable from a knowledge of the scores on the remaining variables. Images and anti-images are used to explain "why" any two variables are correlated with each other, i.e., to reveal the structure of the intercorrelations of the particular universe.

Guttman's analysis takes place within the n dimensional Euclidean test space. This space is defined by the n variables taken as unit vectors with a common origin. The cosine of the angle between any two variables represents the correlation coefficient. The image constitutes a projection of the variable onto the $n - 1$ dimensional space (which is determined by the remaining vectors embedded in the test space). The anti-image is orthogonal to this $n - 1$ space. The multiple correlation coefficient of each variable with each set of the remaining variables is the cosine of the angle between this variable and its respective projection, as well as being the length of the projection.

The image can be conceived as one side of a right-angled triangle. The anti-image is perpendicular to the image vector, and constitutes the distance between the end point of the variable vector and its projection. The hypotenuse of this triangle is the vector being projected.

The square of the multiple correlation (smc) represents the portion of the total variance of the variable that is dependent upon the remaining $n - 1$ variables. By definition, in terms of infinity, the smc constitutes the common variance or communality. It is equal to one minus the unique variance of the variable and is the square of the length of a test vector's projection.

Guttman (1956) recommended the square of the multiple correlation as the "best possible" estimate of communality. These smc 's are well known as Guttman's "universally strongest lower bound."

Rao-Guttman Relationships

Harris (1962b) has developed important relationships between Guttman's image theory and Rao's canonical factor model. He exploited the strengths of both systems to provide an algorithm that replaces judgment by statistical procedures. His solution can be interpreted either as a description of a population *à la* Guttman or as a set of statistically significant factors estimated from a sample.

All critical decisions in analysis are functions of the original data. The one exception is that of judgment in establishing the tolerance limits for the stabilization of the uniqueness estimates.

Harris (1962b) demonstrated that when Guttman's s_j^2 are used as initial estimators of the upper bounds to the unique variances in the Rao procedure, the following obtain: (1) Guttman's "best"

lower bound to the number of common factors is precisely equivalent to the number of roots of $U^{-1}RU^{-1}$ that equal or exceed unity; (2) the characteristic vectors and roots of $U^{-1}RU^{-1}$ can be used to produce both the image covariance matrix and the anti-image matrix; (3) the computing algorithm determines all factors of both matrices only at the end of the first stage of Rao's procedure; (4) the uniqueness estimates are systematically changed in the "proper" direction with each stage of iteration.

In essence, the Harris algorithm permits retaining all the common factors that would be required according to Guttman's assumption (that no sampling error, in the same sense of sampling persons, exists), and it also permits discarding those that are not statistically significant according to Rao's theorem. Thus the factors of the image matrix, common factors that form a Gramian matrix for the set of persons, and canonical factors that satisfy a test of significance can all be secured from the same routine.

Harris' insight into the implications of the Rao-Guttman synthesis led to the development of a "new lower bound" for the unknown communalities. In an effort to find new and more restrictive bounds than s_j^2 , Harris (1963) developed the v_j^2 as initial estimators of U^2 . The Harris v_j^2 indicate a more stringent bound than the well known s_j^2 and seem to provide a more parsimonious solution than Guttman's "best" lower bound as initial estimators of the unique variance.

Classification

The development of multivariate techniques for handling complex data has brought with it various models designed to solve for different end purposes in a parsimonious way.

If an individual chooses to analyze a set of variates only in terms of the data in hand, factoring will take place within the original test space. This procedure will account for the total variance, and factor scores can be computed precisely.

An individual may instead be concerned with obtaining a linear resolution of a set of variates in terms of hypothetical factors. In such a case the analysis would entail projective transformations of the variables out of the original test space into the common-factor space. This mode of analysis provides information about unique variance while focusing on the communalities and the intercorrelations of the variables.

These are the two broad categories that an individual may choose a solution from. Regan (1963) classified these categories as follows. The first is the *component type of analysis*, in which a solution is derived exclusively from the data in hand. The other is the *true factor analysis*, in which a solution is obtained via projective transformations of the data out of the original test space into the common-factor space. Following are the basic models from which specific solutions are constructed.

Component-Type Analysis

Complete component analysis involves a complete factoring of R , the correlation matrix. In this case, W is defined as the set of observed data, scaled and deviated so that WW' yields the matrix of intercorrelations of the observed variables with units in the diagonal, or

$$WW' = R.$$

Linear transformations are made on the original data, i.e., the data are transformed into an uncorrelated derived set of variables. The correlation matrix, R , is extracted and exhausted such that

$$FF' = R$$

which will give as many factors as variables. The matrix F is non-singular, and thus has an inverse F^{-1} . Also $W = FS$, where F is the factor coefficients and S is a matrix of the factor scores for individuals. Consequently,

$$S = F^{-1}W$$

and the factor scores for individuals can be computed exactly.

Incomplete component analysis is equivalent to an incomplete factoring of R , the correlation matrix. Structurally, factoring takes place as described above for complete component analysis. In this case, however, factors that account for a small portion of the variance are "thrown out." Factors that account for the largest variance, in contrast, are retained and are used in reproducing the correlation matrix. The individual must himself decide on the dividing line between "small" and "large" variance.

Image analysis also takes place within the original test space, or at least a subspace of the test space. In this mode of analysis the two matrices of interest include both the image and anti-image matrices. The image matrix is defined by

$$MM' = R + S^2R^{-1}S^2 - 2S^2$$

which is the variance-covariance matrix of the images, with squared multiple correlations of each variable with each of the remaining variables in the main diagonal. S^2 is defined as a diagonal matrix consisting of $1/r^{jj}$. The variance-covariance matrix of the anti-image is defined as

$$AA' = S^2R^{-1}S^2.$$

Since the images are observable, the image factor scores can be determined by a straightforward calculation from the image space. Here,

$$M = (I - S^2R^{-1})Z$$

and $FB = M$, where F is the factor loading matrix and B is the image factor scores. In practice, F will usually be nonsingular, and thus has an inverse. In that case, $B = F^{-1}M$.

Approximation to image analysis is achieved by factoring $R - S^2$, the correlation matrix with the squared multiple correlations in the diagonals. Guttman (1954) demonstrated that the *smc* estimates constitutes his "best" lower bound to the number of common factors. Factoring $R - S^2$ yields not more than $n - 1$ non-negative roots, and at least 1 or $n - m$ imaginary factors resulting from negative roots. $R - S^2$, as a result, is non-Gramian.

This model involves both common and unique factors in the solution of $R - S^2$. The total number of factors exceeds the number of variables, and an inverse does not exist for the factor matrix F . Consequently, factor scores cannot be computed but only estimated, because $FF' = R - S^2$.

Nevertheless, Harris (1962b, 1964) showed that factoring $R - S^2$ is essentially an alternative analysis of MM' , the image covariance matrix. According to his procedure, the first m nonimaginary factors of $R - S^2$ plus the first m estimated factor scores yield a matrix, M^* which is a portion of M . This same M^* is reproduced by the first m factors of the image covariance matrix, together with the first m factor scores.

An approximation to the image covariance matrix can therefore be achieved by putting together the first m factors of $R - S^2$ and the related image factor scores. Since the first m image factor

measurements are embedded in the variable space, these factor scores can be computed precisely instead of merely estimated.

Factor Analysis

Factor analysis implies a communality type of solution in that it is composed of factoring $R - U^2$, where U^2 is a diagonal matrix composed of the unique variance. This mode of analysis postulates a limited number of common factors in an attempt to reproduce the intercorrelations of the set of variables. These constructs, or hypothetical variables, are defined as projections from the variable space into a common-factor space. Consequently, the common-factor solution is always projective in that the common factors alone reproduce the correlation matrix.

The problem confronting an individual estimating the communalities is to find a method that will minimize the rank of the correlation matrix and leave it Gramian. There is no *a priori* knowledge of the communalities; either the rank of the correlation matrix or its diagonal values must be approximated for a factor solution to be obtained. Therefore, the problem in factor analysis lies in finding a diagonal matrix, U , such that $R - U^2$ is Gramian and of minimum rank.

Canonical factor analysis involves the determinantal equation $|R - bU^2| = 0$. This is the Rao type of factor analysis, in which hypothetical variables are considered as projections from the space of the original variables into a common-factor space. In this mode of analysis, the hypothetical measures are uncorrelated with each other, and each measure has a maximum possible correlation with the observed data. The correlation is made maximum by transforming Rao's basic model, $|R - bU^2| = 0$, into $|U^{-1}RU^{-1} - bI| = 0$, and, after Harris' (1962b) operations, it can be defined as $R - U^2 = UQ||b_i - 1||Q'U$. This change translates the communality problem into a problem of rescaling the standardized data by a diagonal matrix, now U^{-1} , to achieve roots, b_i , of $U^{-1}RU^{-1}$, none of which roots are less than unity.

By using the square of the multiple correlations as initial estimates, or, for the sake of convenience, the s_j^2 as estimates of the upper bound for the unique variance, one can obtain a solution which provides maximum information. As an alternative, and possibly a more parsimonious solution, Harris' (1963) v_j^2 can be in-

serted into the diagonal matrix of the original model, here U^2 , as initial estimates of the u_j^2 .

Initial estimates of the U^2 are modified by inserting the estimates into the determinantal equation $|U^{-1}RU^{-1} - b_1| = 0$, then solving for the characteristic roots and vectors. New estimates are constructed. Iterations are continued until the initial estimates reach a stable set of values within the tolerance range selected. Factors can then be constructed such that

$$F = UQ ||b_i - 1||^{\frac{1}{2}}.$$

Harris (1962) shows that the number of factors that will reproduce the "reduced" correlation matrix can be determined by setting m columns for F to be the number of roots, b_i , greater than unity. Rotation procedures will identify those factors that are essentially null.

Canonical factor analysis defines the hypothetical variables as projections into a common-factor space of the variables of the original test space. Factor scores cannot be calculated exactly. In this case, as with all modes of factor analysis, they must be estimated.

Application

Table 1 presents a summary of the two types of analyses: component, including complete, incomplete, image and image approximation, and factor (with u_j^2 and s_j^2 as initial estimates of the canonical form). The data are taken from a study on decision-making by Regan (1963) in which she developed twelve scales to measure positional involvement in policy decisions. These scales or variables are identified in Table 1 as $V_1, V_2 \dots, V_{12}$. The data—observations from 251 adult education administrators in 49 states—are presented as rotated factors by Kaiser's (1958) "normal" varimax procedure.

Complete component analysis entails a complete factoring of R (the matrix of intercorrelations with unities in the diagonal) extracting those latent roots greater than zero. The matrix R is Gramian; therefore, no negative or zero roots exist. The total variance is accounted for by twelve components and is equal to the number of variables, as expected. Table 1 displays only those "factors" for which loadings are greater than .300. Consequently, three common components and four large specific components are

TABLE 1

*Comparison of Component and Factor Types of Analysis
for Twelve Involvement Variables*

Factor	Component Type			Factor Type	
	Complete Component (R)	Incomplete Component (R)	Image	Image Approximation ($R - S^2$)	Canonical (v_i^2) (s_i^2)
Factor I					
V 1	837	834	798	833	862 870
V 2	392	787	672	641	571 552
V 3	846	941	848	917	909 877
V 4	916	823	776	805	832 849
Factor II					
V 5	876	878	819	854	861 870
V 6	951	919	826	859	889 884
V 7	851	885	806	848	871 853
V 8	—	—	—	—	—
Factor III					
V 9	927	902	872	911	935 916
V 10	465	669	642	600	618 632
V 11	901	905	860	912	917 913
V 12	415	777	615	600	604 589
Factor IV					
V 8	957	—	655	692	832 820
V 12	—	—	—	362	322 323
V 7	—	—	—	323	318 307
Factor V					
V 2	896	—	369	457	625 662
Factor VI					
V 12	854	—	—	—	340 450
Factor VII					
V 10	770	—	—	—	339 350
V 7	340	—	336	—	—

Note.—Decimal points have been omitted.

considered meaningful since they account for nearly all the variation.

In contrast to the seven components identified by complete component analysis, an incomplete analysis of R (the intercorrelation matrix with units in the diagonal) was factored by extracting only components for which the latent roots are greater than unity. Thus only three components appear and carry with them most of the variation due to the correlations among variables.

Six factors appearing to have substantial correlations are identified in the complete image analysis. The image variance-covariance matrix is Gramian and has no negative roots; consequently, twelve image factors were extracted, but six of them were trivial and, hence, not meaningful. As is shown in Table 1, both complete com-

ponent and image analyses are very similar in composition. The former, however, identifies two substantial, specific factors which are not clearly separated out by image analysis.

The approximation to image analysis was based on extracting all factors for which the latent roots are greater than zero. As a result, the incomplete factoring of $R - S^2$, which is non-Gramian, yielded five roots that are either zero or negative. Only seven real factors were extracted and, of these, only five have correlations of any significance.

The canonical factor models for the solutions for the twelve involvement variables used v_j^2 and s_j^2 as initial estimators of the unique variance. In both cases the lower boundary to the number of common factors was specified to be the number of roots greater than unity. Tolerance limits were arbitrarily set at .005 in order to converge estimates within a specified range. Both canonical solutions are identical in the number of meaningful factors yielded. The v_j^2 estimates, however, are based on eight factors rather than seven; the eighth, when rotated, proved to be essentially null. In both cases the resultant total involved four common and three specific factors.

In all cases, thus, the largest number of meaningful components or factors was seven, all of which were yielded by the canonical factoring system. Except for complete component analysis (which also yielded the seven, but which presents other problems noted below) all other systems yielded only one portion or another of the total.

Therefore, of all modes of analyses, the canonical form tends to produce the most precise, meaningful factors for the kinds of data presented here. Image analysis, on the other hand, is also very useful, having the advantage of providing precise factor scores for individuals (even though in this study it did not yield as complete data as the canonical form). Depending on the information desired, the use of either system is recommended for these kinds of data. The special advantage of both of these systems is that they are scale free (adaptable either to the correlation or variance-covariance matrix) and thus their uses are almost unlimited.

Two other types listed in the table—complete and incomplete component analyses—are not scale free; they have the weakness that the investigator must choose which components are meaningful

and how many of them are to be retained. In this respect, both allow more room for error than either canonical or image analysis. But like image analysis, they have the advantage of allowing individual factor scores to be computed rather than estimated. A characteristic of both complete and incomplete component analyses is that they allow the investigator to examine only the total variation due to the correlations.

Approximation to image analysis (factoring $R - S^2$) on the other hand, has serious limitations. A complete factoring of $R - S^2$ will always produce negative roots with resultant imaginary factors. Despite its popularity, its problems are insurmountable and, in the opinion of this investigator, its use is not justified now that the properties of complete and incomplete image analyses have been fully investigated.

REFERENCES

- Guttman, L. "Image Theory for the Structure of Quantitative Variates." *Psychometrika*, XVIII (1953), 277-296.
- Guttman, L. "Some Necessary Conditions for Common-Factor Analysis." *Psychometrika*, XIX (1954), 149-162.
- Guttman, L. "'Best Possible' Systematic Estimates of Communalities." *Psychometrika*, XXI (1956), 273-285.
- Harris, C. W. "Some Problems in the Description of Change." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXII (1962), 303-319.
- Harris, C. W. "Some Rao-Guttman Relationships." *Psychometrika*, XXVII (1962b), 247-263.
- Harris, C. W. "Canonical Factor Models for the Description of Change." In: Harris, C. W. (Editor), *Problems in Measuring Change*. University of Wisconsin Press, 1963.
- Harris, C. W. "Some Recent Developments in Factor Analysis." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIV (1964), 193-205.
- Hotelling, H. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology*, XXIV (1933), 417-441.
- Kaiser, H. F. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika*, XXIII (1958), 187-200.
- Lawley, D. N. "The Estimation of Factor Loadings by the Method of Maximum Likelihood." *Proceedings of the Royal Society of Edinburgh*, LX (1940), 64-82.
- Rao, C. R. "Estimation and Tests of Significance in Factor Analysis." *Psychometrika*, XX (1955), 93-111.
- Regan, M. C. "Measurement of Positional Involvement of State Home Economics Leader in Administrative Decisions in Co-

operative Extension." Unpublished Ph.D. dissertation, University of Wisconsin, 1963.

Spearman, C. " 'General Intelligence,' Objectively Determined and Measured." *American Journal of Psychology*, XV (1904), 201-293.

Thurstone, L. L. *Vectors of the Mind*. Chicago: University of Chicago Press, 1935.

RELIABILITY OF COMPOSITE RATINGS¹

JOHN E. OVERALL

The University of Texas Medical Branch
Galveston, Texas

RATING scales are among the most widely used methods of psychiatric and psychological evaluation. In recognition of the fact that ratings by any single rater may include a substantial component of error, a common practice has been to have several raters evaluate each individual and then to average the ratings in the hope of obtaining a more accurate estimate of the "true" score for the individual. Intuitively this makes good sense since the practice is obviously somewhat analogous to increasing the length of a test through addition of items which measure the same thing except for random errors. On the other hand, it is recognized that some raters may be more reliable than others. Intuitively, one feels that it may not always be advantageous to combine with ratings made by a highly experienced and careful rater those made by much less qualified persons. The purpose of this note is to provide a logical basis for answering questions regarding whether or not to combine ratings made by several different raters and, if so, how to combine them in order to maximize reliability.

Methods for computing "optimal" weights have been discussed by several writers (Thurstone, 1931; Mosier, 1943; Green, 1950; Gulliksen, 1950). While these writers have not been concerned specifically with the problem of combining rating made by different observers, the substantive problem involved in combining tests that measure the same thing is essentially the same as the problem of combining ratings of the same thing made by different observers.

¹ This work was accomplished using a computation facility supported by Special Research Resources Grant USPHS 1 PO7 FR-00024-01, NIH.

A major difficulty has been that these writers do not agree on what the optimal solution to the problem is. Without attempting to analyze reasons for the differences in conclusions reached by these different writers, it is obviously difficult to generalize from the test-theory domain to the rating area with any assurance when such a variety of "optimal" solutions have been advanced. Moreover, the results described by several of the test specialists lack the general simplicity required for practical use. A very simple formula for computing optimal weights from estimates of individual rater reliability coefficients is presented in this note.

General Expression for Reliability of Composite

When ratings by several independent raters are combined, a single composite score is obtained which is a linear function of the individual ratings.

$$(1) \quad Y = a_1X_1 + a_2X_2 + \cdots + a_mX_m,$$

where X_1, X_2, \dots, X_m are ratings of the same individual made by m different raters, and a_1, a_2, \dots, a_m are weighting coefficients.

While the most common practice has been to weight equally ratings made by the several raters, any set of a_i values can be used. In most cases, the equal weighting will enhance reliability of evaluation; however, appropriately chosen differential weights can be expected to yield even greater reliability where individual raters are not equally reliable.

Reliability will be defined in the usual way as the ratio of the "true" variance to the "total" variance. (Guilford, 1954, p. 350). The "total" variance of a linear composite, such as equation 1, is a function of the weighting coefficients and of the variances and covariances of ratings by the several raters.

$$(2) \quad \sigma_{\text{TOTAL}}^2 = \sum_{i=1}^m a_i^2 \sigma_{ii}^2 + 2 \sum_{i=1}^m \sum_{j=i+1}^m a_i a_j \sigma_{ij}^2,$$

where a_i ($i = 1, 2, \dots, m$) are the weighting coefficients, σ_{ii}^2 is the variance of ratings by the i^{th} rater, and σ_{ij}^2 is the covariance of ratings made by the i^{th} and j^{th} raters.

The "true" variance of a linear composite, such as equation 1,

assuming rating errors to be random and uncorrelated, is a function of the weighting coefficients, the variances and covariances of ratings by the m raters, and the reliabilities of ratings by the individual raters.

$$(3) \quad \sigma^2_{\text{TRUE}} = \sum_{i=1}^m a_i^2 r_{ii} \sigma_{ii}^2 + 2 \sum_{i=1}^m \sum_{j=i+1}^m a_i a_j \sigma_{ij}^2,$$

where r_{ii} is the reliability coefficient associated with ratings by the i^{th} rater.

It will be noted that equation 3 differs from equation 2 only by the inclusion of the individual rater reliabilities as a factor in the first term on the right. This is true because the variances, σ_{ii}^2 , include both true and error components, while under the assumption of randomness of errors, the covariances reflect only true variation.

A word may be in order here concerning the importance of the assumption of randomness of errors for ratings by different raters. It is well known that raters may have consistent biases, some rate too high and others too low. These average consistent differences affect neither the variances nor covariances among raters since both statistics involve deviations about the mean values of ratings made by the individual raters. Such consistent biases therefore have no effect on the "total" and "true" variances for combined ratings as represented in equations 2 and 3. On the other hand, the tendency for one rater to use a wider range of values—e.g. larger variance of ratings—than another rater uses will affect the values obtained from equations 2 and 3, and this is as it should be.

From the equations given above, the reliability of a linear composite of ratings made by m different raters can be expressed as a function of the weighting coefficients, variances, coveriances and reliabilities of the individual raters.

$$(4) \quad r_{ii} = \frac{\sum_{i=1}^m a_i^2 r_{ii} \sigma_{ii}^2 + 2 \sum_{i=1}^m \sum_{j=i+1}^m a_i a_j \sigma_{ij}^2}{\sum_{i=1}^m a_i^2 \sigma_{ii}^2 + 2 \sum_{i=1}^m \sum_{j=i+1}^m a_i a_j \sigma_{ij}^2}$$

The number of variances entering into the first term in both the numerator and denominator is m , while the number of covariances in the second term is $m(m-1)$. Since the numerator differs from the denominator only by the factor of r_{ii} in the first term, it is

obvious that the reliability of a combination of positively correlated ratings approaches 1.0 as the number of raters is increased because the importance of the second term, which is identical in both the numerator and denominator, increases relative to the first.

Equation 4 is a general expression for the reliability of a linear combination of ratings made by m different judges. Using this equation, one can estimate the reliability of any particular combination of ratings, given estimates of variances, covariances and reliabilities of individual raters. What the equation does, as far as decisions concerning whether or not to combine ratings by several raters are concerned, is to indicate the specific characteristics of the ratings which should be considered in reaching a decision.

It may be noted that if raters can be assumed to be equally reliable and if variances and covariances of their ratings can be assumed to be equal, equation 4 reduces to the familiar Spearman-Brown formula.

$$(5) \quad r_{ii} = \frac{mr_{ii}}{1 + (m - 1)r_{ii}}$$

The purpose here is to consider cases where these assumptions cannot be made.

The Uni-Factor Model for Estimation of Individual Rater Reliabilities

Before considering the problem of selection of weighting coefficients which will tend to maximize the reliability of a composite of ratings made by several different raters, the estimation of individual rater reliabilities should be considered. Frequently one may have *a priori* grounds for assuming differential reliabilities for individual raters. It might be useful at times to consider the question, for example, of whether or not to combine ratings when it is assumed on *a priori* grounds that raters have specified differences in reliability. Frequently, it is possible to obtain an external criterion of the true value of characteristics being rated so that deviations in ratings about these true values can be assessed for each rater. In other situations, the uni-factor model provides a useful means for evaluating individual rater reliabilities.

The uni-factor model is based on the assumption that ratings made by several different raters would correlate perfectly if it were not

for statistically independent errors of measurement. Under this assumption, the covariance between ratings made by two raters is a function of their individual rating standard deviations and individual reliabilities.

$$(6) \quad \sigma_{ii}^2 = \sigma_{ii}\sigma_{jj} \sqrt{r_{ii}} \sqrt{r_{jj}}$$

If it can be assumed that the variances of ratings by different raters are equal, then the covariances among different pairs of raters vary only as a function of differences in individual rater reliabilities. If the ratings have been standardized to unit variance, then the correlation (standard score covariance) between ratings by two different raters is a function of their individual reliabilities.

$$(7) \quad r_{ii} = \sqrt{r_{ii}} \sqrt{r_{jj}}$$

Factor analysis provides a method for estimating individual reliabilities from the matrix of intercorrelations among ratings made by m different raters. In general, the matrix of factor loadings is expected to account for the reduced correlation matrix in the following fashion.

$$(8) \quad R = F'F$$

If we assume that all failures of ratings to correlate perfectly are due to random error, then the diagonal values of the matrix R , as reproduced from the factor loadings, will be the reliability coefficients associated with the individual raters. In familiar terms, the estimated "reliability" for each rater is precisely the "communality" of his ratings as derived from the factor analysis.

Under the assumption that ratings by different raters should correlate perfectly except for random error, only one common factor should be obtained in analysis of the matrix of intercorrelations among ratings by the m raters. What is desired is that single factor which will come closest to accounting for all observed inter-rater correlations. This will be the first principal axes factor of the matrix with communality (reliability) estimates in the principal diagonal. The estimated reliability for each individual rater will then be the square of his loading on the first principal axes factor.

Since the individual rater reliability coefficients are equivalent to communality estimates under the single factor model and all deviations from the model are considered to be random error, the values

used in the diagonal of the matrix which is factor analyzed become critical. An iterative procedure is suggested in which loadings on the first principal axes factor are computed, squares of these loadings inserted into the principal diagonal, and the process is repeated until two successive sets of diagonal values (squared loadings) fail to differ to the third decimal place or beyond. These squared factor loadings will then be considered to represent the best estimates of individual rater reliabilities under the uni-factor model.

It should be noted that the estimation of individual rater reliabilities using factor analysis methods does not require the assumption that only a single common factor is present. A complete set of common factors, if more than one is present, can be extracted. Communality estimates derived from the factor loadings can then be substituted into the correlation matrix, as described previously, and the whole solution can be iterated until the values stabilize. The estimated individual rater reliability is then the sum of squares of factor loading associated with that rater.

Weighting to Maximize Reliability

The reliability of a composite of ratings by several different raters is, in part, a function of the weighting coefficients employed—e.g., the a_1, a_2, \dots, a_m of equation 1. Equal weighting of ratings made by several raters may not result in maximum reliability and, in fact, the equally weighted composite may be less reliable than ratings made by the most reliable single rater. Admittedly, the latter is true only when raters differ rather markedly in individual reliability.

The reliability of a linear composite has been defined as the ratio of "true variance" to "total variance." In matrix notation, this ratio can be expressed as in equation 9.

$$(9) \quad r_{11} = \frac{a' T a}{a' C a},$$

where C is the variance-covariance matrix among ratings by the m different observers, and T is the same variance-covariance matrix except that each diagonal element is multiplied by the individual rater reliability, $r_{11} \sigma_{11}^2$.

Since the "total variance" is the sum of "true" plus "error" variances ($a' C a = a' T a + a' E a$), reliability can be maximized by maximizing the ratio of "true variance" to "error variance."

$$(10) \quad f(r_{ii}) = \frac{a'Ta}{a'Ea},$$

where $\epsilon_{ii} = (1 - r_{ii}) \sigma_{ii}^2$.

This same function can be expressed in terms of the vector z , where $z' = a' E^{1/2}$ and $E^{1/2} E^{1/2} = E$.

$$(11) \quad f(r_{ii}) = \frac{z'E^{-1/2}T E^{-1/2}z}{z'z}.$$

Using the Lagrange method, values of z_i can be computed which will maximize (11) subject to the restriction that $z'z = 1$.

$$(12) \quad (E^{-1/2}T E^{-1/2} - \lambda I)z = 0$$

The values of a needed to maximize the ratio of "true" to "error" variance can be computed readily.

$$(13) \quad a = E^{-1/2}z, \text{ or } a_i = \epsilon_{ii}^{-1/2}z_i \text{ in scalar notation.}$$

It should be noted that in all of these calculations the matrix E is a diagonal matrix; thus $E^{-1/2}$ and $E^{-1/2} T E^{-1/2}$ can be written easily. The general element of E is $\epsilon_{ii} = (1 - r_{ii})$.

A Simplified Formula for Optimal Weights. The general eigen vector solution to the problem of optimal weighting in order to maximize reliability is not new in the area of test theory (Gulliksen, 1950, p. 346). What has not been presented, as far as the present writer can ascertain, is a simplified procedure for calculating optimal weights derived from the more complicated procedure discussed above. Since simplified reliability weighting procedures described by several writers differ rather substantially (Kelley, 1927; Thurstone, 1931; Mosier, 1943), we will restrict our attention to the problem of weighting to maximize the reliability of a composite of ratings made by different raters and we will attempt to develop a procedure which is consistent with the optimal mathematical solution described in the preceding section.

If it can be assumed that ratings by different raters should correlate perfectly except for random errors of measurement, a single common factor will account for T . The general element of T is the product of the square roots of the individual rater reliabilities and variances.

$$(14) \quad t_{ii} = \sqrt{r_{ii}\sigma_{ii}^2} \sqrt{r_{ii}\sigma_{ii}^2}$$

The general element of $E^{-1/2}$ is the reciprocal of the error standard deviation.

$$(15) \quad \epsilon_{ii}^{-1/2} = \frac{1}{\sqrt{(1 - r_{ii})\sigma_{ii}^2}}$$

Going back to equation 12, the general element of the matrix $E^{-1/2} T E^{-1/2}$ is

$$(16) \quad \begin{aligned} \gamma_{ij} &= t_{ij}(\epsilon_{ii}^{-1/2} \epsilon_{jj}^{-1/2}) \\ \gamma_{ij} &= \frac{\sqrt{r_{ii}\sigma_{ii}^2} \sqrt{r_{jj}\sigma_{jj}^2}}{\sqrt{(1 - r_{ii})\sigma_{ii}^2} \sqrt{(1 - r_{jj})\sigma_{jj}^2}} \\ \gamma_{ij} &= \frac{\sqrt{r_{ii}}}{\sqrt{(1 - r_{ii})}} \cdot \frac{\sqrt{r_{jj}}}{\sqrt{(1 - r_{jj})}} \end{aligned}$$

Thus, it is obvious that the elements of z , which will account for $E^{-1/2} T E^{-1/2}$ in equation 12, are proportional to the ratio of square root of *reliability* to square root of *unreliability* under the uni-factor assumption.

$$(17) \quad \lambda^{1/2} z_i = \frac{\sqrt{r_{ii}}}{\sqrt{(1 - r_{ii})}}$$

The elements of a , satisfying equations 9 and 10 (under the uni-factor assumption), are given in (18).

$$(18) \quad \begin{aligned} k a_i &= \epsilon_{ii}^{-1/2} z_i = \frac{1}{\sqrt{(1 - r_{ii})\sigma_{ii}^2}} \cdot \frac{\sqrt{r_{ii}}}{\sqrt{(1 - r_{ii})}} \\ k a_i &= \frac{1}{\sigma_{ii}} \cdot \frac{\sqrt{r_{ii}}}{(1 - r_{ii})}, \end{aligned}$$

where k is an arbitrary constant which can be omitted since it appears in all terms equally.

The weighting coefficients which will maximize reliability of a composite of ratings made by m different observers are a function of individual rater reliabilities and individual rater variances. If variances in ratings by several raters are essentially the same, appropriate weights depend only upon individual rater reliabilities (which assumption yields the weighting coefficients suggested by Thurstone, 1931).

$$(19) \quad a_i = \sqrt{r_{ii}}/(1 - r_{ii})$$

If individual rater reliabilities are essentially equal, but rating vari-

ances differ substantially, the optimal procedure involves weighting by the reciprocal of individual rater standard deviations.

$$(20) \quad a_i = 1/\sigma_{i.}$$

It is interesting to note that the optimal weighting procedure defined by equation 18 involves the two coefficients which have been suggested separately as weighting coefficients by numerous writers. Optimal weighting requires consideration of both individual rater reliability and individual rater variance. If both of these parameters are equal across all raters, equal weighting will be optimal. To the extent that ratings made by different raters tend to be equal in variance and reliability, equal weighting may be quite useful. To the extent that one or both of these parameters tends to vary from rater to rater, differential weighting will prove advantageous.

Example of Optimal Weighting

A psychologist needed to evaluate the level of "anxiety" in each of 100 patients for a research project. Recognizing the low reliability in ratings of "anxiety," he managed to have each patient rated by two psychiatrists (P_1 and P_2), by two nurses (N_1 and N_2), and by two nurse's aids (A_1 and A_2). A question then arose as to whether he should combine ratings made by observers having such very different qualifications.

Intercorrelations among ratings and the standard deviation of ratings made by each observer are presented in Table 1. The correlation between ratings made by P_1 and P_2 is largest, suggesting greatest reliability; however, the "interrater" reliability coefficient computed from ratings by only two observers provides little basis for

TABLE 1
Intercorrelations among Ratings by Six Observers

	P_1	P_2	N_1	N_2	A_1	A_2
P_1	1.00	.55	.41	.38	.32	.30
P_2	.55	1.00	.34	.32	.28	.26
N_1	.41	.34	1.00	.25	.21	.19
N_2	.38	.32	.25	1.00	.19	.20
A_1	.32	.28	.21	.19	1.00	.16
A_2	.30	.26	.19	.20	.16	1.00
$\sigma_{P_1} = 8.0$			$\sigma_{N_1} = 11.1$		$\sigma_{A_1} = 18.2$	
$\sigma_{P_2} = 9.5$			$\sigma_{N_2} = 11.8$		$\sigma_{A_2} = 6.7$	

estimating individual rater reliability. The usual "interrater" reliability is an average (geometric mean) of the individual rater reliabilities where rating errors can be assumed to be random and uncorrelated.

The availability of ratings of "anxiety" made by several different raters permits one to obtain a better estimate of individual rater reliability. This can be achieved by principal axes factor analysis of the matrix of intercorrelations among raters. In the present example, a principal axes analysis, iterated until communality (reliability) estimates stabilized, yielded results presented in Table 2. Loadings on the second and third factors are observed to be quite small so that the uni-factor model appears appropriate. Estimates of individual rater reliability were calculated as the squares of loadings on the first principal factor. It will be noted that the estimated reliability of ratings made by P_1 is different from the estimated reliability of ratings made by P_2 . From Table 2, we see that the estimated reliability of the most reliable single rater is .64.

TABLE 2
*Principal Axes Factor Loadings and Estimates
of Individual Rater Reliability*

	I	II	III	λ^2	r_{ii}
P_1	.798	-.040	.059	.642	.637
P_2	.689	-.113	-.076	.493	.475
N_1	.509	.027	.084	.267	.259
N_2	.484	.134	-.011	.252	.234
A_1	.404	-.008	-.007	.164	.163
A_2	.386	.089	-.077	.162	.149

The next question examined was whether or not the simple equal weighting of ratings made by different observers should be expected to increase reliability in evaluating "anxiety." The matrix of *intercorrelations* among ratings by the six different observers was converted to a *covariance* matrix by pre and post multiplying the correlation matrix by a diagonal matrix containing individual rater standard deviations. Equation 3 was then used to compute the estimated reliability of a composite formed by weighting equally ratings made by the several observers. This was done by setting each of the a_i in equation 3 equal to 1.0. The estimated reliabilities for various combinations of raters are presented in Table 3.

The combination of P_1 and P_2 with equal weighting leads to an

TABLE 3

*Estimated Reliabilities for Various Rater Combinations
Where Equal Weighting is Used*

Rater Combination	r_{tt}
P_1	.637
$P_1 + P_2$.702
$P_1 + P_2 + N_1 + N_2$.685
All raters (P_1, P_2, N_1, N_2, A_1 and A_2)	.666

estimated composite reliability of .70, as contrasted with .64 for the best single rater. Combination of ratings by the two nurses, N_1 and N_2 , with the ratings by P_1 and P_2 leads to an estimated reliability of .69 when equal weighting is employed. Thus, under equal weighting, the addition of ratings by N_1 and N_2 is estimated to decrease the composite reliability as compared with the combination of P_1 and P_2 alone. The equally weighted composite of ratings by all six observers has an estimated reliability of .67. In this example, equal weighting of ratings made by raters with grossly different individual reliabilities does not appear to be an advantageous procedure.

TABLE 4

*Estimated Reliabilities for Various Rater Combinations
Where Optimal Weighting is Used*

Rater Combination	r_{tt}
P_1	.637
$P_1 + P_2$.727
$P_1 + P_2 + N_1 + N_2$.768
All raters (P_1, P_2, N_1, N_2, A_1 and A_2)	.810

Next, the usefulness of differential weighting to increase composite reliability was examined. Optimal weights were computed from individual rater reliabilities (Table 2) and standard deviations (Table 1) using equation 3. The weighting coefficients computed in this way were as follows.

$$\begin{array}{lll}
 P_1 \rightarrow a_1 = .275 & N_1 \rightarrow a_3 = .062 & A_1 \rightarrow a_5 = .027 \\
 P_2 \rightarrow a_2 = .138 & N_2 \rightarrow a_4 = .054 & A_2 \rightarrow a_6 = .068
 \end{array}$$

These differential weighting coefficients inserted into equation 3 yielded the estimates of composite reliability presented in Table 4 for various combinations of raters. In this example, the differential weighting of ratings made by observers having grossly different individual reliability increased estimated composite reliability over the

equally weighted combination and over the most reliable single rater. The estimated reliability of the optimally weighted composite of ratings made by all six observers was found to be .81, while the estimated reliability of a composite derived by equal weighting was only .67. The results suggest that in attempting to evaluate such nebulous constructs as "anxiety," the use of multiple raters and optimal weighting may be advantageous. Simple equal weighting may not be appropriate.

REFERENCES

- Green, B. F., Jr. "A Note on the Calculation of Weights for Maximum Battery Reliability." *Psychometrika*, XV (1950), 57-61.
- Guilford, J. P. *Psychometric Methods*. New York: McGraw-Hill, 1954.
- Gulliksen, H. *Theory of Mental Tests*. New York: Wiley, 1950.
- Kelley, T. L. *Interpretation of Educational Measurements*. New York: World Book Company, 1927.
- Mosier, C. I. "On the Reliability of a Weighted Composite." *Psychometrika*, VIII (1943), 161-168.
- Thurstone, L. L. *The Reliability and Validity of Tests*. Ann Arbor, Michigan: Edwards Brothers, 1931.

A GENERALIZATION OF THE MEDIAN TEST

DALE E. MATTSON
University of Washington¹

THE median test is a nonparametric test which can be used to compare the medians of two or more independent samples. Discussions related to the median test can be found in Brown and Mood (1951), Mood (1950), Moses (1952), and Siegel (1956). The general rationale for the median test may be described as follows. Suppose two or more independent random samples are drawn from populations with a common median. Then except for sampling errors the median of the combined samples ought to be the same as the medians of each of the independent samples. Therefore, the median of the combined samples ought to divide the frequency distribution of each sample approximately in half. A χ^2 test can be used to decide whether the deviations between expected and observed frequencies are greater than would be expected to occur due to random sampling.

The set of procedures used for the median test can be regarded as an example of a more general set of procedures which can be used to compare the entire frequency distributions of two or more independent samples. When the median test is used, the combined median is used to divide the combined distribution into two parts or classes. The decision to use the combined median as a dividing point makes the finding of expected frequencies within the samples quite simple. Within each sample half of the scores are expected to be above the combined median and half below.

It should be apparent that a different percentile value could just as easily be used to divide the combined frequency distribution. If

¹ Now at Division of Educational Research, American Association of Dental Schools, Chicago.

the 25th percentile were chosen, one-fourth of the numbers in each sample could be expected to be below this value. The null hypothesis would then be that the samples are drawn from populations with the same value for the 25th percentile. The decision to use the median as the dividing point is therefore an arbitrary one.

It should also be possible to divide the combined frequency distribution into more than two classes. For example one should be able to test the null hypothesis that two samples are taken from populations with common quartile values. In order to test this hypothesis the first, second, and third quartiles would be computed for the combined distribution thus dividing the distribution into four classes. Expected frequencies within each class for each sample would then be one-fourth of the sample size. The number of classes into which the combined frequency is divided is therefore also arbitrary. In some cases it might be meaningful to let every different score in the combined frequency distribution represent a separate class. The null hypothesis would then be that the samples have been drawn from populations with identical frequency distributions. Of course expected frequencies would be small and the power of the χ^2 test would be reduced.

A general formula for obtaining expected frequencies within the classes for each sample is necessary if dividing points and number of classes are to be decided on an arbitrary basis.

Let f_{ij} represent the frequency in a certain class i for a given sample j . Then the expected frequency for f_{ij} can be obtained from the following formula.

$$f_{ij(e)} = \frac{n_j}{n} \sum_{i=1}^k f_{i1(e)} \quad (1)$$

where

$f_{ij(e)}$ = expected frequency for class interval i for sample j

n_j = size of sample j

n = $\sum n_j$

k = number of samples

and $f_{ij(o)}$ = the observed frequency in class interval i for sample j

As an illustration suppose that two samples of size $n_1 = 30$ and $n_2 = 50$ are available. The combined frequency distribution is divided

into two classes by the combined median. The class interval below the combined median may be designated 1 and the class interval above the combined median as 2. Then by use of equation (1) the expected frequency in interval 1 for sample 1 is obtained as follows.

$$f_{11(*)} = \frac{n_1}{n} \sum_{i=1}^n f_{1i(0)} = \frac{30}{80} (40) = 15$$

In this case $\sum f_{1j} = 40$ since one half of the total number must be below the combined median.

In the example which has been given the use of formula (1) is not necessary. It is apparent that the expected frequency below the combined median in a sample of size 30 is 15. When dividing points are chosen arbitrarily, however, the formula which has been given will be needed in order to obtain expected frequencies.

How does one decide how many classes to use and where to make divisions? These decisions must be made on the basis of the kind of information which is desired. If an experimenter believes that a certain experimental process will raise scores below the 25th percentile and leave other scores unchanged he may wish to use the 25th percentile to divide the combined frequency into two groups. If an experimenter believes that an experimental process will increase high scores, reduce low scores, and leave scores in between unaffected he may wish to divide the combined frequency distribution into three classes with the first and third quartiles as dividing points. In some instances, differences in any part of the distribution may be of interest and there may be as many classes as there are different scores. In this case samples would have to be large enough so that expected frequencies in classes would allow the use of χ^2 (Siegel, 1956, pp. 110).

Example

Suppose that it is possible on a certain test to get scores between 1 and 12 inclusive. During an experiment a group of control Ss and a group of experimental Ss get the scores listed in Table 1. Suppose you wish to test the hypothesis that these scores represent two random samples from populations with the same percentage of scores in the following class intervals: 0 to 3.5, 3.5 to 6.5, 6.5 to 9.5, and 9.5 to 12.5. This hypothesis could be tested as follows.

TABLE 1

Scores	f Control	f Experimental
12	4	4
11	5	8
10	8	11
9	13	10
8	13	8
7	9	7
6	12	6
5	12	7
4	6	6
3	4	9
2	8	12
1	6	12
	$n_c = 100$	$n_e = 100$

TABLE 2

	Control	Experimental
Class 4 9.5-12.5	17	23
Class 3 6.5-9.5	35	25
Class 2 3.5-6.5	30	19
Class 1 0-3.5	18	33

The observed frequencies in each of the four classes for the control group and for the experimental group are shown in Table 2.

The expected frequencies in each class for each sample are easy to find in this case because the samples are of equal size. Formula (1) therefore simplifies to $\frac{1}{2} \sum f_{ij}$ where summation is over all samples. Summing the frequencies in Class 1 over both samples gives a total frequency of 51 for class one. The expected frequency in the control group in class one is therefore 25.5. The same frequency is expected in the experimental group. If samples were of unequal size then expected frequencies would of course be different. Using this same procedure expected frequencies for each group for each class can be obtained. The results have been placed in parenthesis in Table 3.

The computation of χ^2 for Table 3 would be as follows.

$$\chi^2 = \frac{(17 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(35 - 30)^2}{30} + \frac{(25 - 30)^2}{30} \\ + \frac{(30 - 24.5)^2}{24.5} + \frac{(19 - 24.5)^2}{24.5} + \frac{(18 - 25.5)^2}{25.5} + \frac{(33 - 25.5)^2}{25.5}$$

$$\chi^2 = 9.45 \text{ with 3 df}$$

TABLE 3

	Control	Experimental
Class 1	17 (20)	23 (20)
Class 2	35 (30)	25 (30)
Class 3	30 (24.5)	19 (24.5)
Class 4	18 (25.5)	33 (25.5)

Since the probability of χ^2 being this large on the basis of random sampling is less than .02 the null hypothesis would be rejected.

If a median test were used to decide whether these samples are from populations with a common median, the null hypothesis would not be rejected.

Summary

In this paper a generalization of the median test was presented. It was pointed out that the decision to divide a combined frequency distribution into two classes on the basis of the combined median is entirely arbitrary both as to the number of classes and the point of division. In some cases other divisions might be more meaningful. For example it might be meaningful to test the hypothesis that two samples have been drawn from populations with common quartile values, common decile values or common 90th percentiles. A method of obtaining expected frequencies for classes which have been formed arbitrarily was presented. The procedure outlined in this paper was illustrated for a set of fictitious scores.

REFERENCES

- Brown, G. W. and Mood, A. M. "On Median Tests for Linear Hypothesis." *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, California: University of California Press, 1951.
- Mood, A. M. *Introduction to the Theory of Statistics*. New York: McGraw-Hill, 1950.
- Moses, L. E. "Non-parametric Statistics for Psychological Research." *Psychological Bulletin*, XLIX (1952), 122-143.
- Siegel, Sidney. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956.



THE USE OF INFERENCE AS A RESEARCH TOOL

CLIFFORD C. COURSON

Brevard Junior College

DURING the past several decades of psychological inquiry, a number of new and promising ways of viewing and understanding human personality have been presented for scientific investigation. Many researchers and practitioners have accepted or adopted new theoretical constructs and have attempted to explore or evaluate these constructs with instruments and methodologies not entirely consistent with the theoretical frame of reference in which the new constructs were conceived. This confusion between measurement and theoretical frame of reference has been particularly true for "self" psychology (Combs and Soper, 1957) and seems most unfortunate in view of the widespread interest in "self" psychology (Gordon and Combs, 1958). A recent review of research, for example, lists some 493 references of studies on the self concept (Wylie, 1961). A closer inspection of these references reveals that most are actually studies of the self-report, a construct which has been shown to have no significant relationship to the self concept (Combs, Soper, and Courson, 1963).

It has been suggested that measurement of the self concept or related theoretical constructs can be achieved through the use of inferences by trained observers (Combs, 1958).

Yet there are many who would deny the use of inference as a research tool for the behavioral sciences. Other scientific disciplines have long recognized the value of "going beyond the data," attempting to infer the direction, the pattern and the meaning of events.

The use of inference as research data is often distressing to psychologists accustomed to working in the more orthodox external

frame of reference. They miss the apparent solidity of external measurements. The use of inference in research is not new, however. The physical sciences have been using it effectively for many years. Through inference the physicist is able to deal with electricity or atomic particles which no one has ever seen and the astronomer can locate stars outside the range of his telescope. Without the use of inference many of our great advances in physical science could never have been made. Inferential research can be highly accurate and controlled when it is clearly understood and properly handled (Combs and Soper, 1963a).

In brief, inference as a research tool has not been extensively used by research workers in the behavioral sciences. While many teachers, clinicians and other practitioners commonly use inference as a technique for gathering operational data, formal research has largely confined itself to external and more orthodox methods for data collection.

The few researches which have used inference as a technique for gathering data have so far failed to fully or clearly demonstrate the reliability of inference as a research tool. For example, a study involving inferences about the perceptual organization of counselors reported percentages of scoring agreements among the raters who made the inferences (Combs and Soper, 1963b). In a more refined study of child perceptions, the estimates for the reliability of inferences were based upon an evaluation of the communality figures resulting from a factor analysis of the inferred ratings (Combs and Soper, 1963a). While this post-analysis of reliability is statistically valid, it also serves to amplify the need for a direct and more conventional study of the reliability of inference.

Problem

The heart of the issue would seem to lie in whether or not inference as a research tool can be shown to be a reliable method for collection of data. Any science is, after all, based upon observation and the accuracy of observation. In the behavioral sciences, the accuracy of observation is usually called the *reliability* of a measurement. The *validity* of inferential data was not evaluated in the present study, but several researches report significant relationships between inferred data and external criteria (Combs and Soper, 1963a; Combs and Soper, 1963b; Courson, 1963; Fink, 1962; Lamy,

1962; LaVerd, 1961; Malpass, 1953). The present study, then, was primarily concerned with testing whether inference as a research tool can be a reliable methodology for the collection of data for the behavioral sciences.

The study was designed to measure the reliability of inferences in two ways. The first of these ways involved testing whether several trained observers working independently could make significantly similar inferences from common samples of simulated behavior. It was felt that this procedure would reflect the reliability of inferences *between* several observers.

Secondly, the study was designed to test whether trained observers could make significantly similar inferences from a simulated behavior sample when the behavior sample was re-rated after a period of approximately one month. It was felt that this would serve as a check on the reliability *within* each rater or, in other words, a check on the internal consistency for each rater.

The hypotheses for the study were stated as predictions as follows: (1) *Several trained observers working independently will make significantly similar inferences about the same sample of simulated behavior and,* (2) *the same observers will make significantly similar inferences when samples of simulated behavior are re-rated after a period of approximately one month.*

The Projective Essays

The behavior samples collected for this study of inference consisted of pairs of projective essays written by each of 64 subjects who were seniors at the P. K. Yonge Laboratory School of the University of Florida. The essays were required of the subjects as a part of their twelfth-grade writing laboratory and approximately fifty minutes was allowed for completion of each of the essays.

1. The Assigned Projective Essay—The first behavior sample collected from each subject in the study consisted of an assigned essay written on the topic "A Teenager's Advice to the World." It was felt that this would be a non-threatening way of asking subjects how things looked to them, how they saw the world around them. Some pilot experimentation by the writer with a small number of subjects had demonstrated that this was a promising source of information about the internal perceptual organization of subjects. In fact, the pilot work revealed that fairly valid inferences could be

made from random samples of written work from high school students.

2. The Thematic Apperception Test Essay—The second behavior sample collected consisted of an essay written in response to selected pictures from the Thematic Apperception Test. Three cards, number 1, number 4, and number 20, were simultaneously presented to the subjects with the instruction to select one card and write an imaginative essay about the card.

The final simulated behavior sample from which the inferences in this study were made consisted of a pair of projective essays as described above. The essays were collected by the writing laboratory teachers and delivered to a typist. The essays were then typed, identifying information was removed, and each pair of essays was assigned a code number such that no cues about the identity of the subjects were available. The contents of the pairs of typed essays constituted the simulated behavior samples evaluated by the raters in this study.

The Observers

In order to evaluate the reliability of inference as a research tool, two observers, in addition to the writer, were selected and trained for this study. Prior to this study, the writer had several years' research and practical experience in making systematic inferences from behavior observations. This prior experience has been in connection with other research projects carried out at the University of Florida. The trained observers for the study were a counseling psychologist and an advanced doctoral student in educational psychology, both at the University of Florida.¹

The training of the observers took place in the framework of several informal sessions devoted to discussion of pertinent theoretical and practical implications of concern to the study. In order to make perceptual inferences, the raters were encouraged to use themselves as instruments, to use any and all data from the observations that had some personal meaning. The raters were encouraged to make the kind of inferences they would make for their own uses, inferences which would not necessarily need to be supported or de-

¹ The writer is grateful for the assistance of Dr. James Lister and John Benton. This study would not have been possible without their generous effort and cooperation.

fended by concrete evidence. They were instructed to use the full resources of their experiences as behavioral scientists and as sensitive human beings in the making of inferences.

The Perceptual Factors Rating Scale

The specific variables selected for measurement in this study of inference were certain perceptual factors identified by Combs (1962) as factors underlying or giving rise to the adequate personality. These perceptual factors were an essentially positive view of self, a feeling of wide identification with others, and an openness to experience. In addition, a summation score for the three perceptual factors was included as a fourth variable in the study.

A simple, nine-point rating scale was constructed for use by the raters in the study. The scale served to anchor or quantify the inferences made by raters about each subject in the study for the four perceptual variables under consideration. A facsimile of the perceptual rating form used in the study for recording inferences is reproduced below.

PERCEPTUAL FACTORS RATING SCALE

Code # _____	Rater _____							
1. How does this person see himself?								
1	2	3	4	5	6	7	8	9
Essentially Negative							Essentially Positive	
2. To what extent is this person identified with others?								
1	2	3	4	5	6	7	8	9
Strongly Alienated							Strongly Identified	
3. To what extent is this person open to his experience?								
1	2	3	4	5	6	7	8	9
Closed Prejudiced Much distortion							Essentially Open and Accepting	
Total _____								

The three raters in the study worked independently in making inferences based upon the projective essays. Since the procedures for making inferences demanded both time and sensitivity on the part of the raters, a rule was established that not more than one rating would be completed in a one-hour period of time. It was felt that this procedure would help to minimize or eliminate rater "sets."

Results, Discussion, and Implications

The writer scored the essay sets for all of the 64 subjects in the study. The 64 essay sets were then randomly divided into two groups and each of the trained raters scored 32 sets. In order to check the reliability of the scoring procedures for the inferences, Pearson correlation coefficients were computed between the scores inferred by the writer and the scores inferred by each of the two trained raters. The results of these computations are listed in Table 1 with levels of significance as determined by a *t*-test (McNemar, 1955).

TABLE 1
*Pearson Correlations between Writer and Two Raters
for Perceptual Factors Ratings*

Perceptual Factor	Rater #1 Pearson r ($N = 32$)	Rater #2 Pearson r ($N = 32$)	Mean r
1. How does this person see himself?	.46***	.38*	.42**
2. To what extent is this person identified with others?	.42**	.43**	.42**
3. To what extent is this person open to his experience	.55***	.51***	.53***
4. Total score for Perceptual Factors	.53***	.48***	.51***

* = significant beyond .025 level of confidence, one-tailed *t*-test

** = significant beyond .01 level of confidence, one-tailed *t*-test

*** = significant beyond .005 level of confidence, one-tailed *t*-test

The results listed in Table 1 indicate that the degree of agreement between the scoring of the writer and the independent scoring of the two trained raters had little possibility of occurring by chance. In order to achieve a summary picture of the reliability between the writer and the two trained raters, the Pearson correlations obtained were converted to *z*-scores and averaged. The resulting average correlations for each perceptual factor are also listed in Table 1. The analysis of reliability between raters demonstrated significant agreement for all of the inferred perceptual factors. The magnitude of the correlations for agreement were in line with expectations based upon consideration of the limited behavior sample studied and the nature of the concepts involved. In view of the foregoing, the reliability of

the scoring procedures between raters was considered to be adequately demonstrated.

A second method for evaluating the reliability of inferences involved a repetition of the ratings after the elapse of a period of time. This was, in essence, an evaluation of the reliability *within* each rater, a check on the internal consistency of the raters.

Approximately one month after completion of the evaluation of reliability of inferences between raters, the writer and each of the two trained raters rerated ten randomly selected essay sets from the original group of 64. The scores inferred in the original ratings were compared with the scores on the second set of inferred ratings for each rater by means of a Pearson correlation. The results of the analysis of reliability within raters are presented in Table 2.

TABLE 2
Pearson Correlations for Evaluation of Reliability within Three Raters

Perceptual Factors	Writer N = 10	Rater #1 N = 10	Rater #2 N = 10
1. How does this person see himself?	.72**	.74**	.80***
2. To what extent is this person identified with others?	.78***	.83***	.79***
3. To what extent is this person open to his experience?	.81***	.79***	.62*
4. Total score for perceptual factors	.84***	.75**	.84***

* = significant beyond .05 level of confidence, one-tailed t-test

** = significant beyond .01 level of confidence, one-tailed t-test

*** = significant beyond .005 level of confidence, one-tailed t-test

The reliability analysis presented in Table 2 showed that a consistently high and statistically significant consistency was demonstrated within each rater in the study. The implication was that the raters tended to make very similar inferences when rating the same sample of simulated behavior after a lapse in time. In other words, regardless of the validity of the inferred ratings in the study, the internal consistency of each rater was demonstrated to be statistically significant.

The results of the study, then, lend support to the position that inferred ratings by trained observers may be reliable data. The findings demonstrated that several observers working independently made significantly similar inferences about the same samples of

simulated behavior. In addition, the same observers made significantly similar inferences when they re-rated samples of simulated behavior after a period of approximately one month.

The author does not wish to imply that inference as a research tool represents a panacea for measurement in the behavioral sciences. The use of inference does, however, offer new avenues for research and certainly presents a new kind of data supplementing conventional methodologies. The implication is that many of the data we have about human behavior do not readily lend themselves to orthodox or external analysis. Further analysis of data already available, using inference as a research tool, could presumably lead to a number of new insights. In much of our research we have, perhaps, been focusing upon objective, observable behaviors without fully understanding the real meaning and purposes of behavior. Our reading of high school essays, for example, is usually rather superficial and does not include a conscious attempt to understand the writer's basic purposes, how he sees himself and the world around him. In brief, it seems quite probable that we have been ignoring data which could lead to some valuable insights into human personality.

The making of inferences about the personal meanings of behavior offers new diagnostic insights. The inferential question is not so much directed at the external components of behavior; but rather, the question is directed at the internal dynamics, the general purposes and directions of behavior. The focus is upon environmental and life conditions *as they are seen by the individual*, as they "exist" in the perceptual field of the individual.

Finally, if we would more fully understand human personality, we must "go beyond the data" or be willing to settle for a collection of isolated facts. The present study has demonstrated that inferences about internal conditions based upon observations of external behaviors can be a reliable methodology. The implication of this finding is that we should give some attention to better understanding the data already on hand rather than concentrating our efforts on seeking new data.

Summary

This study was designed to evaluate the reliability of inference as a research tool. Three observers independently rated samples of

simulated behavior in the form of written projective essays and made inferences about certain internal factors in the perceptual organization of 64 subjects. A significant degree of agreement was found between the inferred ratings of the three observers. These same observers rerated the simulated behavior of ten randomly selected subjects approximately one month after the original observations. A significant degree of agreement was found between the original and the repeated inferred ratings. Hence, the reliability of inference as a research tool was demonstrated *between* observers and *within* observers.

The writer briefly discussed some of the implications of these findings for measurement in the behavioral sciences.

REFERENCES

- Combs, A. W. "A Perceptual View of the Adequate Personality." *Perceiving, Behaving, Becoming*. Edited by A. W. Combs, Washington: Association for Supervision and Curriculum Development, 1962 Yearbook, 50-64.
- Combs, A. W. "New Horizons in Field Research: The Self Concept." *Educational Leadership*, XV (1958), 315-319. 328.
- Combs, A. W. and Soper, D. W. *The Relationship of Early Child Perceptions to Achievement and Behavior in the Early School Years*. Gainesville, Florida: University of Florida, U. S. Office of Education Cooperative Research Project No. 814, 1963. (a)
- Combs, A. W. and Soper, D. W. "The Perceptual Organization of Effective Counselors." *Journal of Counseling Psychology*, X (1963), 222-226. (b)
- Combs, A. W. and Soper, D. W. "The Self, Its Derivative Terms, and Research." *Journal of Individual Psychology*, XIII (1957), 134-145.
- Combs, A. W., Soper, D. W. and Courson, C. C. "The Measurement of Self Concept and Self Report." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIII (1963), 493-500.
- Courson, C. C. *The Relationship of Certain Perceptual Factors to Adequacy*. Unpublished doctoral dissertation, University of Florida, 1963.
- Fink, M. B. "Self-Concept as it Relates to Academic Underachievement." *California Journal of Educational Research*, XIII (1962), 57-62.
- Gordon, I. J. and Combs, A. W. "The Learner: Self and Perception." *Review of Educational Research*, XXVIII (1958), 433-444.
- Lamy, Mary W. *Relationship of Self-Perceptions of Early Primary Children to Achievement in Reading*. Unpublished doctoral dissertation, University of Florida, 1962.
- LaVerd, John. *The Relationship of the Achievement Motives as*

- Projected into TAT Stories to Children's School Achievement.*
Unpublished doctoral dissertation, Utah State University, 1961.
- Malpass, L. F. "Some Relationships between Students' Perceptions of School and Their Achievement." *Journal of Educational Psychology*, IVIV (1953), 475-482.
- McNemar, Quinn. *Psychological Statistics*. Second Edition, New York: John Wiley and Sons, 1955.
- Wylie, Ruth C. *The Self Concept*. Nebraska: University of Nebraska Press, 1961.

A FACTOR-ANALYTIC STUDY OF SPONTANEOUS-FLEXIBILITY MEASURES¹

FRANK B. MAY AND ALAN W. METCALF

Washington State University

THE purposes of this study were (1) to determine the extent to which spontaneous flexibility is independent of certain changes in test instructions and scoring procedures, and (2) to compare the effectiveness of three different means of deriving scores of spontaneous flexibility.

Background and Rationale

The phenomenon called "flexibility" has been investigated for many years (Cattell and Tiner, 1949; Guetzkow, 1951; Guilford, Frick, Christensen, and Merrifield, 1957; Kleemeier and Dudek, 1950; Luchins, 1942; Oliver and Ferguson, 1951). More recently, however, flexibility has become a popular variable in creative thinking studies. Both Torrance (1960, 1962) and Guilford (1952, 1956, 1961), for example, have utilized flexibility measures to a great extent in determining the creative potential of adults and children.

So far, however, Torrance and Guilford have differed considerably on how to measure flexibility. The measurements of the two investigators differ in important respects such as scoring procedures, mental sets induced by test instructions, and examination tasks. Furthermore, Guilford's studies indicate at least two types of flexibility—spontaneous and adaptive—whereas Torrance makes no attempt to separate the two types. The effect of fluency on Guilford's measurements of flexibility has been made evident through factor

¹This study was part of the Spokane Dropout Study financed by the State Department of Public Instruction in Olympia, Washington, and directed by Lloyd B. Urdal, Department of Education, Washington State University.

analysis. The effect of fluency on Torrance's measurements of flexibility has not been made evident. Moreover, the effect of different test instructions on measurements of flexibility has not been systematically studied, although Torrance (1962-b) has examined such effects on measurements of originality. In another study of originality, Christensen, Guilford, and Wilson (1957) found that tests on which the subjects were instructed to be original and tests on which the subjects were *not* instructed to be original both had high loadings on the same factor. It would seem equally important to ascertain whether flexibility is similarly independent of the mental set induced by test instructions.

In addition to further psychometric delineation of flexibility, scoring procedures which are both conservative of time and reasonably objective need to be established. Torrance and Guilford have developed at least three different flexibility scores for tests of similar content. With Guilford's "Brick Uses" test, a score for spontaneous flexibility is derived by counting the different *categories* of uses employed by the subject—categories such as building uses, uses as a weight, and uses as a pounding instrument. With Torrance's "Tin Can Uses" and "Book Uses" a multi-category scheme is also used (although the measurement is called "flexibility" rather than "spontaneous flexibility"). A multi-category scheme, however, can be very time consuming and quite subjective.

In addition to the category scheme for scoring "Brick Uses," Wilson and other associates of Guilford have derived another score for spontaneous flexibility by counting the number of unusual or *alternate uses* which the subject thinks of for an object. Each item in their Alternate Uses Test (1960) presents the name of a common object, along with a statement of its most common use. The subject is asked to think of six other uses which are less common. The test consists of nine items in sets of three, with four minutes given for each set. A sample item is presented to the subject. The score for this test is based on the number of acceptable responses; "vague," "impossible," and "overworked" responses are considered unacceptable. This test eliminates some of the shortcomings of tests which must be scored by the category scheme. However, other shortcomings seem evident: (1) Limiting a subject to six responses per item may stifle his flexibility "just when he gets going." (2) Informing the subject of the common use of the object and of the necessity to

consider less common uses seems, logically, to provide too much of a clue to the subject; to measure *spontaneous* flexibility, it appears illogical to *tell* the subject to be flexible. (3) To measure *spontaneous* flexibility it seems inappropriate to show the subject *how* to be flexible by providing him with a sample item. (4) The scoring protocol is somewhat lacking in specificity.

In addition to the category scheme for scoring "Tin Can Uses" and "Book Uses," Torrance has derived another score for flexibility by counting the number of different *principles* employed by a subject in thinking of improvements for an object. Torrance has delineated 20 principles, including maximization, minification, and addition, for this scoring procedure. Essentially, this is another category scheme. However, no matter what object the subject is asked to "improve," the same twenty principles can be used to score his responses for flexibility; a new set of categories does not have to be established for each object. This scoring procedure shares the deficiency of the category scheme in that scorers frequently have difficulty classifying responses. (A more severe weakness will be discussed later.)

Modified Testing and Scoring Procedures

In an attempt to overcome some of the limitations in the testing and scoring procedures which have been discussed, modified procedures for testing and scoring spontaneous flexibility were developed by the writers. Our purpose was to develop procedures which would meet the following criteria:

1. Lack of clues to the subjects that spontaneous flexibility is called for.
2. Relatively fast and objective scoring.
3. Relatively low correlation with fluency.

Following is a test item and scoring procedure designed to meet those criteria:

Instructions:

List as many uses as you can think of for one or more pencils.

You will have only three minutes.

Scoring Protocol (unconventional uses):

Score only one point for *marking* with the lead (writing, shading, drawing, etc.); only one point for *erasing* (no matter

what is erased); and an additional point for each other use, other than marking or erasing. *However,*

- (a) Score no point for an impossible use, but avoid a strict interpretation of "impossible." (An example of an "impossible" use would be "a cheap tool for cutting soft diamonds.")
- (b) Score no more than two points for an action verb repeated within a single item (e.g., "poking people," "poking cats.")
- (c) Score no additional point for any response which *can* be placed in a "conventional use" category, (e.g., "in arithmetic class" can be placed in the category of "marking.")
- (d) Score no point for any response which is not an actual use (e.g., a preparation for use, such as "sharpening;" an action which is only related to the object, such as "something to forget;" dispensation after use, such as "putting in your pocket.")

Method

Subjects

The subjects of the present study consisted of 332 eighth-grade students in a Spokane, Washington high school. The students were heterogeneous with respect to scholastic abilities, sex, and socioeconomic status.

Tests

The following group tests (which included only verbal stimuli) were administered to the students:

Creative Thinking Battery

1. Table Fork Improvement. This test was adapted from Torrance's "Fire Truck Improvement" (1960) and scored for fluency (total number of responses), and also for spontaneous flexibility (Torrance's principles). A mental set of fluency was encouraged by warning the subjects that they had only three minutes to list as many improvements as they could.
2. Chalkboard Improvement. (Same as number one.)
3. Pencil Uses. This test was adapted from Guilford's "Brick Uses"

(1952) and scored for fluency and also for spontaneous flexibility (unconventional uses). A mental set of fluency was encouraged by warning the subjects that they had only three minutes to list as many uses as they could.

4. Tin Can Uses (Torrance, 1960). This test was scored for fluency and also for spontaneous flexibility (Torrance's categories). A mental set of fluency was encouraged as in number three.
5. Broom Uses. (Same as number three.)
6. Book Uses. (Same as number four.)
7. Unusual Substitutes. This test, consisting of two parts, was adapted from Wilson's "Unusual Uses" (1953) and was scored for uncommonness of response. The score was derived by weighting each response from one to five depending on the frequency of its occurrence in this population. This test was used as a reference variable for "conceptual adaptive flexibility" (Guilford, 1957) and as a buffer between the fluency set and the subsequent flexibility set. The test instructions follow:

In the following test list five things which might be substituted for each object below. In other words, if you did not have the following objects available, what might you use instead? Try to think of the most *unusual* substitutes that you can. However, they must be reasonable and practical, even if they are unusual.

A mental set of flexibility was further encouraged by telling the subjects to "Take your time. Don't hurry. You will have a full seven minutes."

8. Alarm Clock Improvement. This test was adapted from Torrance's "Fire Truck Improvement" (1960) and scored for spontaneous flexibility (Torrance's principles). A mental set of flexibility was encouraged by telling the subjects to "list many *different kinds of improvements*. For example, besides improving the *sound* of an alarm clock, also list other kinds of improvements. Take your time. Don't hurry. You will have a full five minutes."
9. Rocking Chair Improvement. (Same as number eight, except no example was given.)
10. Table Knife Uses. This test was adapted from Guilford's "Brick Uses" and was scored for spontaneous flexibility (unconventional uses). A mental set of flexibility was encouraged by tell-

ing the subjects to "list many *different kinds* of uses. For example, besides listing cutting uses, also list other kinds of uses. . . . Take your time. Don't hurry. You will have a full five minutes."

11. Needle Uses. (Same as number ten, except no example was given.)

Differential Aptitude Tests

Only the score on the "Verbal Reasoning Test" was used in this study. "Verbal Reasoning" is a power test consisting of 50 verbal-analogy items. It was designed to measure a combination of the "verbal ability" and "deductive reasoning" factors (Carroll, 1959) and correlates highly with intelligence scores (Frederiksen, 1959). Both the first and last components of each analogy are omitted, and the subject must choose from several alternatives the words which will complete the analogy. Because the population used in this study was heterogeneous, the Verbal Reasoning Test was included as a reference variable for verbal intelligence.

Treatment of Data

To increase the range of the scores and the reliability of the inter-correlations, several pairs of scores were combined, resulting in 11 variables as shown in Table 1. A perusal of variables 4-7 will indicate how two of the scoring procedures for spontaneous flexibility were each paired with two mental sets. An obvious weakness in our design is the lack of a partner variable for variable 8. It would have been desirable to administer another pair of "Uses" tests under a mental set of flexibility, which then would have been scored for spontaneous flexibility, using the category scheme. The omission of this variable was partly a concession to expediency: a maximum of 50 minutes was allotted for the creative thinking battery; also, scoring time would have been increased considerably. The omission was also a result of theoretical considerations: Torrance had developed categories for only two tests—"Tin Can Uses" and "Book Uses;" the writers considered it appropriate to test only *his* categories and *his* principles, since Torrance seems to consider both schemes as equivalent means of scoring flexibility. The effect of this omission of a partner variable for variable 8 will be discussed in a later section.

TABLE 1

Means, Standard Deviations and Reliability Estimates of Eleven Tests Administered to 332 Heterogeneous Eighth Graders

Var.	Tests	Scoring	Set	Task	Mean	S. D.	r_{11} *	R **
Creative Thinking Tests								
1	1 + 2	Fluency	Flu.	Improvements	8.56	3.55	.80	.99
2	3 + 5	Fluency	Flu.	Uses	13.29	5.56	.74	.99
3	4 + 6	Fluency	Flu.	Uses	16.61	6.78	.62	.99
4	1 + 2	Principles	Flu.	Improvements	6.68	2.46	.62	.88
5	8 + 9	Principles	Flex.	Improvements	8.07	2.94	.77	.87
6	3 + 5	Unconv. Uses	Flu.	Uses	6.68	3.37	.50	.93
7	10 + 11	Unconv. Uses	Flex.	Uses	9.97	5.01	.77	.98
8	4 + 6	Categories	Flu.	Uses	9.72	3.74	.51	.90
9	7 a	Uncommonness	Flex.	Substitutes	12.58	3.63	.23	N
10	7 b	Uncommonness	Flex.	Substitutes	12.60	4.58	.22	N
Differential Aptitude Test								
11	V. R.	Correct Responses		Analogies	20.29	8.03	.88	N

* Estimated for variables 1-8 by a stepped-up correlation between test scores. Estimated for variables 9 and 10 by using the communality as a low estimate. Estimate for variable 11 by using the parallel-forms coefficient reported in the test manual.

** Interjudge reliability

N Not computed

The 11 variables were factor analyzed by means of the CDC 1604 electronic computer, using a program designed by Harris (1962). This program utilizes the principle axes algorithm, with the roots and vectors obtained from the matrix $U^{-1}RU^{-1}$ where U is a diagonal matrix of uniqueness estimates and R is the correlation matrix. U^2 is estimated by one minus the square of the multiple correlation coefficient. Factors corresponding to eigenvalues greater than 1.0 are retained. These factors are subjected to a varimax rotation.

Results

Table 2 shows the intercorrelations of the eleven variables described in Table 1. Tables 3 and 4 show the unrotated and rotated factor matrices. Although six factors were extracted and rotated, only four will be described in Table 5. The other two factors displayed loadings which the writers considered psychologically insignificant (see Table 4), and accounted for less than 3 per cent of the total variance. The four factors described in Table 5 accounted, respectively, for 18%, 16%, 10%, and 7% of the total variance.

Discussion

It is likely that FACTOR I is similar to the dimension referred to by Guilford (1952) as "ideational fluency." However, since variables

TABLE 2*
*Intercorrelations of Eleven Variables***

	1	2	3	4	5	6	7	8	9	10
2	54									
3	50	76								
4	85	44	44							
5	54	44	46	52						
6	30	51	31	27	24					
7	17	32	32	18	45	28				
8	37	47	55	36	35	60	32			
9	18	18	12	17	20	18	25	15		
10	16	19	21	16	17	27	18	29	25	
11	02	09	00	10	08	27	16	22	10	13

* Decimal points omitted

** See Table 1 for description of variables

TABLE 3*
*Unrotated Factor Matrix for Eleven Variables***

Variable	Factors					
	I	II	III	IV	V	VI
1	-.83	-.36	.00	.05	-.02	-.01
2	-.76	.31	.20	.06	-.13	-.09
3	-.74	.29	.32	-.01	.07	.03
4	-.79	-.41	-.09	.02	.02	.02
5	-.64	-.04	-.03	-.33	.05	-.04
6	-.50	.39	-.36	.18	-.07	.05
7	-.37	.28	-.14	-.41	.00	-.09
8	-.61	.38	-.22	.09	.18	.11
9	-.24	.06	-.20	-.22	-.20	-.07
10	-.27	.18	-.22	-.07	-.03	.16
11	-.13	.17	-.37	.01	-.02	-.15

* Decimal points omitted

** See Table 1 for description of variables

TABLE 4*
*Rotated Factor Matrix for Eleven Variables***

Variable	Factors					
	I	II	III	IV	V	VI
1	-.85	.29	-.09	-.10	-.04	.02
2	-.31	.75	-.21	-.19	-.12	.00
3	-.30	.77	-.07	-.17	.09	.10
4	-.86	.19	-.12	-.11	.00	.06
5	-.47	.28	-.06	-.46	.10	.07
6	-.14	.33	-.64	-.11	-.07	.17
7	-.06	.23	-.19	-.55	.08	.06
8	-.21	.44	-.52	-.13	.20	.22
9	-.13	.03	-.17	-.35	-.16	.03
10	-.09	.12	-.27	-.18	.00	.25
11	-.01	-.04	-.41	-.16	-.01	-.07

* Decimal points omitted

** See Table 1 for description of variables

1 and 4 were not experimentally independent (the same pair of tests were scored for two different things), it is difficult to know whether Factor I involves "fluency with improvements" or "flexibility with principles." On the other hand, since Table 4 shows that several fluency measures had moderate to high correlations with this factor, and since Table 5 shows that variable 5 also had a high loading on the adaptive flexibility dimension, it seems reasonable to hypothesize that Factor I represents "fluency with improvements" rather than "flexibility with principles." Most important for this study, it appears that the "principles" scheme yields scores which are highly contaminated by the relative fluency of the subjects. Even

TABLE 5
Four Factors Extracted from Intercorrelations of Eleven Variables

Variable	Scoring	Task	Mental Set	Loading
FACTOR I: Ideational Fluency (A)				
4	Principles	Improvements 1 + 2	Fluency	.86°
1	Fluency	Improvements 1 + 2	Fluency	.85°
5	Principles	Improvements 8 + 9	Flexibility	.47°
FACTOR II: Ideational Fluency (B)				
3	Fluency	Uses 4 + 6	Fluency	.77°
2	Fluency	Uses 3 + 5	Fluency	.75°
8	Categories	Uses 4 + 6	Fluency	.44
FACTOR III: Spontaneous Flexibility				
6	Unconv. Uses	Uses 3 + 5	Fluency	.64°
8	Categories	Uses 4 + 6	Fluency	.52°
11		Verbal Reasoning		.41°
FACTOR IV: Adaptive Flexibility				
7	Unconv. Uses	Uses 10 + 11	Flexibility	.55°
5	Principles	Improvements 8 + 9	Flexibility	.46
9	Uncommonness	Substitutions 7 a	Flexibility	.35°

* Highest loading for this variable on any of the four factors

variable 5, which was independent of variable 1 in testing sequence, mental set, and scoring, had its highest loading on Factor I.

FACTOR II also appears similar to the dimension which Guilford calls "ideational fluency." It is interesting, however, that with this population two different types of ideational fluency were isolated. It appears that Factor II is a dimension which relates primarily to the task of thinking up uses for objects under the mental set of fluency. It is tempting, then, to label Factor II as "ideational fluency: uses" and to label Factor I as "ideational fluency: improvements." These labels, of course, are hunches. It is significant, for purposes of this investigation, that variable 8 (scored for categories), had a high loading on this factor, whereas variable 6 (scored

for unconventional uses), had a relatively low one (see Table 4). This suggests that "unconventional uses" as a scoring procedure was less affected by fluency than "categories" as a scoring procedure.

A comparison of FACTOR III with factors isolated by Guilford (1952, 1957, 1961) suggests that this factor is similar to the dimension which Guilford calls "spontaneous flexibility." This factor appears to be unrelated to ideational fluency and seems, rather, to be the result of the varying degree to which the subjects spontaneously shifted in their thinking from one type of use to other types of uses for an object. Both the "unconventional uses" scheme and the "categories" scheme were effective means of scoring this dimension. The "principles" scheme seems to be an ineffective means of scoring spontaneous flexibility. The high loading of the verbal reasoning score on this factor is to be expected. Several of the items in the Verbal Reasoning Test seem to require flexibility, since they contain words which have alternative meanings and require the subject to shift in his thinking in order to use the meaning which permits interpretation leading to a correct answer. These items are somewhat similar to those employed by Guilford's test called "Implied Uses" (1952), which had a high loading on the factor that Guilford labeled "spontaneous flexibility." Verbal Reasoning also may have had a high loading on Factor III because of its high correlation with general intelligence. It has been shown by Ripple and May (1962) that scores of spontaneous flexibility have a high correlation with intelligence scores when the population tested is heterogeneous.

FACTOR IV is labeled "adaptive flexibility" for two reasons: First, variable 9, a reference variable, is similar to those variables in Guilford's investigations (1952, 1956) which have been indicators of "originality." Guilford (1957) has more recently suggested that this type of "originality" would be more appropriately described as "conceptual adaptive flexibility." Second, for all three variables with a high loading on this factor, the mental set was flexibility. Since the mental set was flexibility rather than fluency, Factor IV definitely should not be considered as *spontaneous* flexibility. Both the "unconventional uses" scheme and the "principles" scheme were effective means of scoring this dimension. It seems quite possible that "categories" would have been an effective means of scoring this dimension also, if additional "uses" tests had been administered under a mental set of flexibility and scored for different categories.

Summary and Conclusions

The effects of different test instructions and scoring procedures on the dimension called "spontaneous flexibility" were examined by means of factor analysis. Three means of scoring spontaneous flexibility were compared to determine their relative effectiveness. Analysis was based on the test scores of a heterogeneous group of 332 eighth-grade students.

Within the limits of this study, the following conclusions are offered:

1. Spontaneous flexibility is probably *not* a factor which is independent of scoring system, mental set induced by test instructions, and examination task.
2. Spontaneous flexibility probably should be measured with tests of uses rather than improvements, under a mental set of fluency rather than flexibility.
3. Spontaneous flexibility probably should be scored either by the "unconventional uses" scheme or the "categories" scheme, but not by the "principles" scheme.
4. Adaptive flexibility can be measured either with tests of uses or tests of improvements, as long as the mental set is flexibility.
5. Adaptive flexibility can be scored by the "unconventional uses" scheme or the "principles" scheme or possibly the "categories" scheme.
6. The "unconventional uses" scheme appears to be the most economical and objective means of deriving a fluency-free score of spontaneous flexibility or adaptive flexibility.

REFERENCES

- Carroll, J. B. in *The Fifth Mental Measurements Yearbook*, edited by Oscar K. Buros, Highland Park, New Jersey: Gryphon Press, 1959.
- Cattell, R. B. and Tiner, L. G. "The Varieties of Structural Rigidity." *Journal of Personality*, XVII (1949), 321-341.
- Christensen, P. R., Guilford, J. P., and Wilson, R. C. "Relations of Creative Responses to Working Time and Instructions." *Journal of Experimental Psychology*, LIII (1957), 82-88.
- Frederiksen, N. in *The Fifth Mental Measurements Yearbook*, edited by Oscar K. Buros, Highland Park, New Jersey: Gryphon Press, 1959.
- Guetzkow, H. "An Analysis of the Operation of Set in Problem-

- Solving Behavior." *Journal of General Psychology*, XLV (1951), 219-244.
- Guilford, J. P. "Creative Thinking Abilities of Ninth-Grade Students." A paper read at the convention of the American Educational Research Association in Chicago, February 24, 1961.
- Guilford, J. P., Frick, J. W., Christensen, P. R., and Merrifield, P. R. "A Factor-Analytic Study of Flexibility in Thinking." Reports from the Psychological Laboratory No. 18; University of Southern California, April, 1957.
- Guilford, J. P., Kettner, N. W., and Christensen, P. R. "A Factor-Analytic Study Across the Domains of Reasoning, Creativity, and Evaluation: II. Administration of Tests and Analysis of Results." Reports from the Psychological Laboratory No. 16; University of Southern California, March, 1956.
- Guilford, J. P., Wilson, R. C., and Christensen, P. R. "A Factor-Analytic Study of Creative Thinking: II. Administration of Tests and Analysis of Results." Reports from the Psychological Laboratory No. 8; University of Southern California, July, 1952.
- Harris, C. W. "Some Rao-Guttman Relationships." *Psychometrika*, XXVII (1962), 247-263.
- Kleemeier, R. W. and Dudek, F. J. "A Factorial Investigation of Flexibility." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, X (1950), 107-118.
- Luchins, A. S. "Mechanization in Problem Solving, The Effect of Einstellung." *Psychological Monographs*, LIV (1942), No. 6 (Whole No. 238).
- Oliver, J. A. and Ferguson, G. A. "A Factorial Study of Tests of Rigidity." *Canadian Journal of Psychology*, V (1951), 49-59.
- Ripple, R. E. and May, F. B. "Caution in Comparing Creativity and IQ." *Psychological Reports*, X (1962), 229-230.
- Torrance, E. P. "Creative Thinking of Children." *Journal of Teacher Education*, XIII (1962), 448-460.
- Torrance, E. P. "Testing and Creative Talent." *Educational Leadership*, XX (1962), 7-10+.
- Torrance, E. P., Yamamoto, K., Schenetzki, D., Palamutlu, N., and Luther, B. *Assessing The Creative Thinking Abilities of Children*. Bureau of Educational Research, College of Education, University of Minnesota, 1960.
- Wilson, R. C., Christensen, P. R., Merrifield, P. R., and Guilford, J. P. "Alternate Uses: Form A: Manual of Administration, Scoring and Interpretation (Preliminary Edition)." Beverly Hills, California: Sheridan Supply Company, 1960.
- Wilson, R. C., Guilford, J. P., and Christensen, P. R., "The Measurement of Individual Differences in Originality." *Psychological Bulletin*, (1953), 362-370.

EXPERIENCE, EXPERTNESS, AND IDEAL TEACHING RELATIONSHIPS¹

WILLARD E. REITZ²

The Menninger Foundation

PHILLIP S. VERY

University of Rhode Island

AND GEORGE M. GUTHRIE

Pennsylvania State University

A number of writers have been interested recently in the essential ingredients of "good" human relationships. Rogers (1958), for example, attempted to define the *general characteristics* of such relationships. He suggested that common characteristics are found across human helping relationships such as therapy, counseling, supervision, and teaching.

Heine (1950) compared descriptions of factors in therapeutic experiences prepared by 24 subjects who had been treated by therapists of three divergent schools. He found some factors were common to all schools and that these factors were described as the primary therapeutic agent.

Fiedler (1950) reported a study in which a 75-item Q sort was used to describe the ideal therapeutic relationship. He had a number of therapists of different theoretical viewpoints and different degrees of competence describe this ideal relationship. He found that

¹ This research was completed while the first two authors were U. S. Public Health Service Predoctoral Research Fellows at the Pennsylvania State University. The authors wish to thank Drs. Gardner Murphy and Riley Gardner for critically reading the manuscript. Thanks also are due to Miss Lolafaye Coyne for statistical advice and to Dr. Hugh S. Brown for his assistance in planning the study.

² Now at The University of Western Ontario, London, Canada.

- “1. Therapists of different schools do not differ in describing their concept of an ideal therapeutic relationship.
2. The ability to describe this concept is probably a function of expertness rather than theoretical allegiance.
3. Nontherapists can describe the ideal relationship in the same manner and about as well as therapists. The therapeutic relationship may therefore be but a variation of good interpersonal relationships in general.”

Soper and Combs (1962) noted that most of the research to date dealt primarily with common factors in the therapeutic relationship. Following Fiedler (1950) they posited that some of the common elements may extend to other professional helping relationships, specifically to the teacher-student relationship. These investigators administered Fiedler's Q-sort items in a form suitable for the teaching situation to “good” elementary and secondary teachers. They found a correlation of .81 between their good teachers and the therapists' composite sort. This correlation compared favorably with highest individual intercorrelations of Fiedler's therapists. Soper and Combs concluded: “. . . our hypothesis that teachers' ratings of an ideal teacher-student relationship are highly similar to the therapists' ideal therapeutic relationship, is amply supported by the data.” Unlike Fiedler, Soper and Combs did not use a novice or inexperienced group of teachers, and they were not concerned with theoretical allegiance.

The present study seeks to extend the question, originally raised by Fiedler, into higher education. Again, however, the problem of theoretical allegiance will not be dealt with. It would seem that theoretical allegiance in the area of teaching in higher education is much less well defined than it is in psychotherapy. It is conceivable, of course, that college teachers could be classified in terms of their preference for directive versus nondirective teaching methods, (Cantor, 1946), the lecture, tutorial, or recitation methods (Highet, 1950) or similar classifications. In the present study, the question is asked whether university teachers of varying content areas and different lengths of teaching experience will describe the ideal teaching relationship differently.

It may be expected that since content areas are vastly different in higher education, that the technique, methods, and interpersonal relationships may vary concomitantly. The other possibility is that

good interpersonal relationships will not vary across content areas, or that an essentially single, unified concept of an ideal teaching relationship will emerge.

A related issue concerns the relationship of the teaching situation to psychotherapy. If Fiedler as well as Soper and Combs are correct, the ideal teaching relationship should be described by teachers in a manner similar to the way in which therapists describe the ideal therapeutic relationship. The alternative possibility here is that the therapeutic and teaching relationships have little in common. If this latter possibility is so, then low or negative relationships should result.

Method

Teachers were asked to describe by means of a 75-item Q sort the undergraduate teacher-student relationship which they considered ideal. Respondents were selected from six colleges within the Pennsylvania State University: Liberal Arts, Education, Agriculture, Engineering, Home Economics, and Chemistry-Physics. Two rather gross levels of teaching experience were utilized. A dean from each of the six colleges was asked to submit names of five or six teachers. The criteria to be used in this selection were as follows: (1) The teachers should be quite experienced in terms of years of undergraduate teaching, (2) they should currently be teaching undergraduates, (3) they should be considered "good" teachers as reflected by student and faculty opinion. The deans were informed of the nature and purpose of the study. They were further informed that teachers would be selected at random from the list of names until two teachers from each college agreed to participate.

These teachers then constituted the "experienced" group. A "novice" group of teachers was selected by obtaining from the deans a list of graduate students in their respective colleges who were just beginning to teach. The requirements stipulated here were: (1) that they should not have taught before, (2) that they should be doing some kind of undergraduate teaching currently for the first time. In the case of Education majors, it was found necessary to permit some previous teaching, that is, student teaching. Even with this exception, however, the entire novice group were newcomers to undergraduate instruction at the college level. As with the experienced group, two novices were selected from each college. The 12 experi-

enced teachers had a total of 205 years teaching experience, with a mean of 17.08 years. Only one teacher refused to participate. One further teacher was dropped because of the time involved in having the Q sort returned. Both of these were replaced in the sample. All of the beginning teachers, however, participated.

The measuring instrument was the same Q-sort statements used by Soper and Combs (1962) who simply changed Fiedler's original items by substituting the word *teacher* for *therapist* and the word *student* for *patient* whenever they occurred. These 75 statements were typed on slips of paper, shuffled, and given to each subject with instructions to sort them into seven categories, with 1, 7, 18, 23, 18, 7, and 1 statements in each category, with the items most characteristic of their idea of an ideal teaching relationship at one extreme and the statements least characteristic of this same relationship at the other extreme. In accordance with Stephenson's technique, each statement was assigned a score from one to seven, depending on the category it had been placed in by the teacher. The composite scores of each teacher were then correlated with the composite score of every other teacher, and the resulting correlation matrix was factor analyzed by the principal components method.

Results

The scores earned by all teachers correlated positively with each other, ranging from .24 to .81 with a median of .57. The correlations and factor loadings are shown in Table 1.

The scores earned by experienced teachers correlated more highly

TABLE 1
Product Moment Intercorrelations* and Factor I Loadings

		1	2	3	4	5	6	7	8	9	10	11	12
Liberal Arts	<i>E</i> ^b	1											
	<i>E</i>	2	.77										
	<i>I</i> ^c	3	.58	.65									
	<i>I</i>	4	.51	.60	.59								
Education	<i>E</i>	5	.56	.68	.58	.54							
	<i>E</i>	6	.55	.65	.62	.54	.66						
	<i>I</i>	7	.58	.75	.61	.54	.54	.65					
	<i>I</i>	8	.58	.64	.55	.81	.54	.55	.58				
Agriculture	<i>E</i>	9	.62	.71	.60	.53	.68	.66	.59	.57			
	<i>E</i>	10	.56	.68	.57	.60	.72	.63	.62	.68	.72		
	<i>I</i>	11	.51	.68	.66	.66	.52	.61	.56	.64	.57	.69	
	<i>I</i>	12	.65	.71	.54	.55	.58	.64	.55	.61	.65	.51	.55

TABLE 1 (Continued)

			1	2	3	4	5	6	7	8	9	10	11	12
Engineering	<i>E</i>	13	65	75	65	56	62	67	71	65	66	63	55	69
	<i>E</i>	14	50	69	50	56	55	51	63	58	52	56	65	54
	<i>I</i>	15	61	64	59	55	54	60	53	63	58	65	68	71
	<i>I</i>	16	53	61	59	56	53	61	54	66	59	65	61	65
Home	<i>E</i>	17	30	33	43	46	24	35	32	46	33	39	52	49
Economics	<i>E</i>	18	62	70	53	60	59	58	60	65	61	64	54	72
	<i>I</i>	19	58	64	52	65	61	68	65	70	67	67	60	67
	<i>I</i>	20	47	47	32	43	45	43	51	46	44	44	40	57
Chemistry- Physics	<i>E</i>	21	56	67	56	55	75	72	55	61	71	68	54	66
	<i>E</i>	22	60	75	66	65	66	72	70	60	73	68	67	66
	<i>I</i>	23	46	55	51	49	45	45	41	59	45	51	59	50
	<i>I</i>	24	55	72	56	56	68	63	63	60	65	69	55	56
Factor I			75	87	75	76	77	79	77	80	80	82	78	81
Loadings														

* For one r $p < .05$; for all other r 's $p < .01$.

^a Experienced.

^b Inexperienced.

TABLE 1 (Continued)

		13	14	15	16	17	18	19	20	21	22	23	24
<i>E</i>	13												
<i>E</i>	14	65											
<i>I</i>	15	56	60										
<i>I</i>	16	62	53	71									
<i>E</i>	17	36	37	45	45								
<i>E</i>	18	72	60	61	65	49							
<i>I</i>	19	66	58	61	69	34	67						
<i>I</i>	20	65	57	43	35	35	58	48					
<i>E</i>	21	70	55	65	63	36	69	61	47				
<i>E</i>	22	73	64	63	65	40	64	63	55	79			
<i>I</i>	23	58	54	56	60	43	56	58	45	52	53		
<i>I</i>	24	67	57	49	51	27	57	66	43	67	65	54	
		85	75	79	78	52	82	82	62	82	86	69	80

with each other across colleges than did the scores earned by experienced teachers of a given college with novices within that same college. To take an example, the average intercorrelation, (an average of Fisher z 's), of experienced-engineering teachers' scores with all other experienced teachers' scores was higher than was the average intercorrelation of experienced-engineering teachers' scores with novice-engineering teachers' scores. This pattern held consistently across all six colleges. These results are statistically significant ($p < .05$) using the Fisher Exact Probability Test (Siegel, 1956).

Table 2 presents these relationships. The reason for the rather gross disparity between scores of Home Economics teachers and teachers from the other colleges is not apparent. An inspection of their individual ratings of items yielded no clear-cut answer.

TABLE 2

Average Intercorrelations (Transformed to Fisher z 's) of Experienced Teachers of Any Given College with, (1) All Other Experienced Teachers and (2) Inexperienced Teachers Within Same College

	Average z of Experienced Teachers of a College with All Other Experienced Teachers	Average z of Experienced Teachers of a College with Inexperienced Teachers of Same College
Liberal Arts	.72	.67
Education	.72	.65
Agriculture	.74	.71
Engineering	.71	.66
Home Economics	.56	.55
Chemistry-Physics	.79	.69

A second finding of interest is the fact that the scores of experienced teachers correlated more highly among themselves than did the scores of inexperienced teachers among themselves. The average z score was significantly higher for the experienced teacher-experienced teacher intercorrelations ($z = .720$) than for the inexperienced teacher-inexperienced teacher ($z = .653$), ($t = 2.21$, $p < .05$).

The correlation matrix of the 24 teachers was factor analyzed by the principal components method. The first factor extracted accounted for 60.42 percent of the total variance. The remaining four factors extracted accounted for 5.08%, 3.98%, 3.21%, and 2.92% of the variance respectively. The first five factors, then, account for 75.61 percent of the total variance which estimates the amount of common variance. Eighty percent of this common variance is accounted for by the first factor. Thus it appears that there is one major factor. No test was done to determine the number of significant factors. The secondary factors appear to be small enough to be of negligible importance relative to the first, general factor.

Since only one factor was found, the ratings could be pooled and a combined rating of the ideal obtained. The three correlations found between (a) Fiedler's composite sort scores of the ideal therapeutic relationship, (b) the composite sort (mean ratings) of the experienced teachers, and (c) the mean ratings of the novice teachers are

shown in Table 3. As can be seen, the scores earned by experienced teachers correlate more highly with composite sort scores earned by Fiedler's therapists than do the scores earned by inexperienced teachers. At the same time, however, the scores earned by novice teachers correlate quite highly with the scores of experienced teachers. All of the correlations in Table 3 are significantly different from each other at the .01 level or beyond.

TABLE 3

Intercorrelations among Therapists' Composite Sort, and Experienced and Inexperienced Teachers' Ratings

	A	B	C
A. Fiedler's Composite Sort			
B. Experienced Teachers' Ratings	.75		
C. Inexperienced Teachers' Ratings	.61	.91	

One way of comparing the composite sorts for the different groups across studies is to list the eight most and eight least characteristic statements for each of the groups. These statements, representing those in categories 1, 2, 6, and 7 are given in Table 4.

In examining the eight statements classified as most characteristic by teachers and the eight considered most characteristic by Fiedler's experts, Soper and Combs (1962) noted that four of the eight were common to both. In the present study only two were common to all

TABLE 4

Most and Least Characteristic Q-Sort Items

Items Common to all Four Groups	
<i>Most Characteristic</i>	
The teacher really tries to understand the student's feelings.	
The teacher is well able to understand the student's feelings.	
<i>Least Characteristic</i>	
The teacher is hostile toward the student.	
The teacher feels disgusted by the student.	
The teacher is very unpleasant to the student.	
The teacher's own needs completely interfere with his understanding of the student.	
Items Common to Three Groups	
<i>Most Characteristic</i>	
The teacher gives and takes in the situations (E, I, S-C)*.	
The teacher sees the student as a co-worker on a common problem (E, F, S-C).	
The teacher usually maintains rapport with the student (E, I, S, S-C).	

TABLE 4 (Continued)

Least Characteristic

- The teacher is rejecting to the student (*E, I, S-C*).
- The teacher is punitive (*E, F, S-C*).
- The teacher shows no comprehension of the feelings the student is trying to communicate (*F, I, S-C*).

Items Common to Two Groups

Most Characteristic

- The teacher is pleasant to the student (*E, I*).
- The teacher is able to participate completely in the student's communication (*F, S-C*).
- The teacher directs and guides the student (*I, S-C*).
- The teacher greatly encourages and reassures the student (*I, S-C*).

Least Characteristic

- The teacher is very condescending to the student (*E, I*).
- The teacher cannot maintain rapport with the student (*E, S-C*).
- The teacher acts in a very superior manner toward the student (*E, I, S*).

Unique to One Group

Most Characteristic

- The teacher maintains a friendly, neutral attitude throughout (*E*).
- The teacher reacts with some understanding of the student's feelings (*E*).
- The teacher's comments are always right in line with what the student is trying to say (*F*).
- The teacher always follows the student's line of thought (*F*).
- The teacher's tone of voice conveys the complete ability to share the student's feelings (*F*).
- The teacher treats the student as an equal (*F*).
- The teacher acts neither superior nor submissive to the student (*I*).

Least Characteristic

- The teacher talks down to the student as if he were a child (*E*).

* *E* = experienced teachers of the present study, *I* = inexperienced teachers of the present study, *F* = Fiedler's composite sort, and *S-C* = Soper and Combs' ratings.

four groups. If the inexperienced teachers are eliminated from consideration, however, thus keeping the comparison to "experts," three statements are common to the three remaining groups. These statements deal with understanding the one helped and seeing him as a co-worker on a common problem.

Of the eight statements selected as least characteristic of an ideal relationship by Fiedler's therapists and Soper and Combs' teachers, seven were common to both. Again adding only the experienced teachers of the present study, this number is reduced to five. It would appear that the ideal elementary and secondary teaching relationship is more akin to conditions of ideal therapy than is the ideal undergraduate teaching relationship.

A number of variables could have influenced these results, for example, the therapeutic relationship deals, in the main, with only one other individual. Soper and Combs' teachers taught elementary and secondary students at a university laboratory school. While the number of students per class in this school is not known, it is unlikely that it matched the average number of students per class of the university teachers of the present report. From this point of view, the "helper-helped" ratio may be an important consideration. If so, one might expect that a different kind of institution of higher education, say a small liberal arts college, would be expected to yield results more similar to Fiedler's.

Discussion

In the present study, an attempt was made to answer the following questions:

1. Are "good" experienced teachers' descriptions of the ideal teacher-student relationship similar to therapists' descriptions of an ideal therapeutic relationship?
2. Does content, and therefore to some extent technique, affect the teachers' beliefs about the nature of the ideal teaching relationship?
3. Does length of teaching experience affect teachers' description of the ideal relationship?

The results of the study indicate that the content area does not affect the relationship which the teacher strives to create with his students. To the extent that the selection criteria can be considered valid, the correlations suggest that expertness and experience influence the type of relationship the teacher sets as a goal. One might expect that at least some of the wide content and technique differences which exist, for example, between the Colleges of Education and Chemistry-Physics, would have been reflected in the sortings obtained from these two colleges. No such difference appeared. The findings suggest that teachers generally agree on the ideal teaching relationship, although experienced teachers agree more, both with themselves and with therapists. The fact that only one factor was found indicates that both experienced and inexperienced teachers had essentially the same ideal in mind. This conclusion is further borne out when the first factor loadings are considered. The differences between factor loadings are not striking although they tend to be higher for experienced teachers.

While the amount of variance accounted for by the first factor is large, it is necessary to ask about the extent to which this result may be due to the nature of the items. That is, items stating extreme points of view may be sorted approximately the same ways by all teachers. This would inflate the correlations and thus tend to generate a manifold first factor. If more items were included, for example, on which good professors disagree, the first factor would have accounted for a smaller portion of the total variance, and the second and third factors would have been larger (Goodling and Guthrie, 1956). It is necessary to keep this point in mind in evaluating the results.

The present results tell us nothing about whether or not a layman could describe the ideal teaching relationship as well as teachers. This question, however, needs attention.

A different but related problem concerns the question of whether the ability to *describe* an ideal relationship is related to the ability to actually *develop* one. It is certainly possible that a discrepancy could exist here. Still another problem concerns the students' conception of the ideal teaching relationship. Are there discrepancies between the students' ideal and the teachers' ideal? If so, in what way? The Q-sort technique would seem to make many of those problems susceptible to empirical investigation.

REFERENCES

- Cantor, N. *The Dynamics of Learning*. Buffalo: Foster and Stewart, 1946.
- Fiedler, F. E. "The Concept of an Ideal Therapeutic Relationship." *Journal of Consulting Psychology*, XIV (1950), 239-245.
- Goodling, R. A. and Guthrie, G. M. "Some Practical Considerations in Q-sort Item Selection." *Journal of Counseling Psychology*, III (1956), 70-72.
- Heine, R. W. "A Comparison of Patients' Reports on Psychotherapeutic Experience with Psychoanalytic, Nondirective and Adlerian Therapists." Unpublished doctoral dissertation, University of Chicago, 1950.
- Hight, G. *The Art of Teaching*. New York: Knopf, 1950.
- Rogers, C. R. "The Characteristics of a Helping Relationship." *Personnel and Guidance Journal*, XXXVII (1958), 6-17.
- Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956.
- Soper, D. W. and Combs, A. W. "The Helping Relationship as Seen by Teachers and Therapists." *Journal of Consulting Psychology*, XXVI (1962), 288.

EVALUATIVE RESPONSES TO AFFECTIVELY POSITIVE AND NEGATIVE FACIAL PHOTOGRAPHS: FACTOR STRUCTURE AND CONSTRUCT VALIDITY¹

CARROLL E. IZARD AND JUM C. NUNNALLY
Vanderbilt University

For a number of years we have investigated the role of emotion, or affect, in various aspects of behavior. We have found two distinct approaches to be fruitful. The first involves experimental induction of affect and the measurement of subsequent effects on behavior ratings, intellectual productivity (Wehmer and Izard, 1962), stereoscopic selective perception (Izard and Jennings, 1963), learning, perceptual threshold, and preference (Izard, et al., 1963). These affect-induction studies have shown that the quality of affect (positive, negative) expressed by *E* made substantial differences in *Ss*' perception and performance.

The second approach involves use of emotionally toned stimuli, particularly pictures of the human face, to evoke responses that might be used in assessing aspects of the perceiver's personality. An early study along this line (Beier, Izard, Smock, and Tougas, 1953; 1957) obtained a simple like-dislike rating on each of a fairly large series of pictures. It showed that male and female *Ss* responded differentially to age and sex categories of pictures. In another study (Izard, 1953; 1959) *Ss* were requested to give a brief report of their first impressions of pictures and to make a preference ranking of

¹ This report is based on part of a program of research supported by contract between Group Psychology Branch, Office of Naval Research, and Vanderbilt University.

Special thanks to Dr. Howard Rolf and the staff of the Vanderbilt Computer Center for assistance in programming the statistical analyses. Programming and computing were partially supported by National Science Foundation Grant NSF-G1008.

A summary of this paper was presented at APA convention, 1962.

the picture stimuli. In this study favorableness of feeling toward pictures of people was greatly different for normal Ss and paranoid schizophrenics. Analysis of the verbal responses showed wide intra-individual discrepancies in responses to pictures manifesting positive and negative affect. In view of these and other findings, we felt it would be fruitful to make a more detailed study of responses to facial photographs that have been judged to be positive and negative emotional stimuli.

For convenience, we shall term our device for studying responses to emotionally toned facial photographs the First Impression Rating Scale (FIRS), but we want to emphasize that the FIRS is considered a research technique, not a formal test.

The FIRS was designed, both by way of instructions and in terms of response alternatives permitted, to elicit trait judgments of facial photographs based on immediate feelings and impressions. It was expected that these perceptual-affective responses would measure attitude and feeling toward people in general. The purpose of this article is to summarize the results of a series of investigations concerning the psychometric properties and construct validity of these responses.

The Stimulus Materials

The Pictures

Several hundred pictures from magazines and college annuals were inspected to find pictures which varied in attractiveness and which did not represent obvious stereotypes. These pictures were culled to a group of fifty pictures. Judges were then asked to sort the pictures into two groups, those which evoked positive feelings and those which evoked negative feelings. Twelve pictures unanimously judged by ten people as negative and twelve unanimously perceived as positive constituted the final series of stimulus pictures. The 24 pictures were divided into two sets (Forms I and II), each containing six positive and six negative pictures. The two sets were approximately equal in attractiveness, according to the judges' preference ratings.

Only pictures of males were utilized. The people pictured had an estimated age range of 20 to 65. The original pictures were reproduced on standard 2" \times 2" black-and-white transparent slides.

The Polar Adjective Rating Scales

In considering adjectives which were to be used in rating the pictures, the aim was to select a wide variety of polar pairs representing many different facets of personality. A second requirement was that one adjective in each pair be more positive than the other. Each pair of polar adjectives constituted the end-points of a 9-point rating scale. The more positive adjective was on the high end of the scale; thus, the higher the numerical rating, the more favorable the judgment. Fifteen such scales were applied to the pictures of Form I, and an approximately synonymous set of polar adjective pairs were utilized in rating the pictures of Form II. The technique yielded three scores: FIRS T, the total score or algebraic sum of all ratings to all pictures; FIRS+, the algebraic sum of the ratings to the six positive pictures, and FIRS—, the algebraic sum of the ratings given to the six negative pictures.

The Testing Technique

Using a slide projector, Ss were tested in groups. Before the pictures were exposed the following instructions were given:

Whenever we see a person for the first time we usually develop some impressions of him. Just from the first meeting we get an idea about the kind of person he is—whether he is likable, interesting or stimulating. I have here the pictures of some people. I'd like to see how they impress you—what you think about them, how you feel about them—when you see them for the first time. You will have about a minute and one-half to make your evaluations of each picture. Be sure to base your ratings on what *you* think and how *you* feel about the person.

Each picture in the series of 12 was exposed for about 90 seconds. While the picture was projected on the screen the subject rated it on each of the 15 polar-adjective scales.

Reliability

The Form I-Form II reliability for the FIRS T ($N = 100$ males) was .85. In evaluating the Form I-Form II reliability, it should be remembered that these are independent alternate forms consisting of different pictures and different polar adjective scales. Test-retest

reliability for Form I ($N = 34$ females) was .81. The interval between the two administrations was one week.

Factor Analysis of Pictures

It was hoped that an analysis of correlations among responses to the twelve pictures would show whether we had been correct in summing over all pictures to obtain a total score. Also, such an analysis might indicate subgroupings of the pictures which should be scored separately. The results in these regards were so straightforward that complex methods of factor analysis and rotation were not needed.

The first step in the analysis was the intercorrelation (product-moment) of individual differences in total ratings of each picture by the sample of 326 arts and science (A&S) men on Form I. An inspection of the correlation matrix showed rather clearly the nature of the groupings involved. First, all correlations among pictures were positive, the average correlation being .22. This provided some justification for our practice of adding over all twelve pictures to obtain scores for individuals. Second, positive pictures tended to go together, negative pictures tended to go together, and the two types of pictures have relatively little in common. The average correlation among positive pictures was .38, the average correlation among negative pictures was .30, and the average correlation between the two types of pictures was only .12. This was taken as strong evidence that there are two major clusters, or factors, among the pictures.

To further demonstrate the tendency for positive and negative pictures to subsume different factors, a group centroid analysis was made. A centroid was placed among the positive pictures, and a second centroid was placed among the negative pictures. Used in this way the group centroid method simply was a means for correlating all of the pictures with a sum of scores on a subgroup of the pictures. The results are shown in Table 1.

The analysis was repeated for the A&S men on Form II, which also are shown on Table 1. Not shown in the table are identical analyses performed on the A&S females and on the engineering students.

In all analyses, the same fact stands out: the positive and negative pictures subsume largely different factors, which we take as

TABLE 1

*Factor Loadings for the Twelve Pictures of Form I and Form II
Determined by the Group Centroid Method
N = 326*

	Picture No.	Form I		Form II	
		F_1	F_2	F_1	F_2
Negative Pictures	3	.20	.65	.25	.70
	5	.20	.67	.21	.70
	6	.14	.64	.25	.66
	7	.21	.57	.36	.67
	9	.11	.70	.22	.71
	10	.15	.65	.41	.72
Positive Pictures	1	.72	.03	.75	.23
	2	.64	.26	.64	.40
	4	.64	.31	.68	.26
	8	.73	.11	.74	.17
	11	.74	.20	.74	.28
	12	.69	.17	.58	.41

the major finding of this study. Representative of the results for the two factors on the two forms and on the three samples are those for A&S men on Form I. The correlation between the two factors was .26. The mean sum of squared loadings (variance explained) for factors 1 and 2 were .26 and .23, respectively. When the correlation between factor 1 and factor 2 was partialled out of the variance explained by factor 2, the net per cent of variance explained by the two factors was .39.

Of course, in factor analysis, when one has 12 tests (here pictures) one ends up with 12 factors, if one wants to entertain tiny loadings and/or to deal with factors that say very little about the phenomena. For two reasons, in the present analysis, we think that two factors are enough. First, the common variance is almost gone after the variance attributable to the two factors is extracted from the correlation matrix. In the analysis of responses to Form I by A&S men, after the variance attributable to the two factors is removed, the average (ignoring signs) residual coefficient (excluding diagonal elements) was only .084. Also, an inspection of these residuals indicated no other factor that was either statistically strong or scientifically interesting.

Internal consistency reliability estimates for the two factors and two forms are given in Table 2. Separate coefficients are shown for three groups of Ss.

TABLE 2

Reliability Estimates (Coefficient Alpha) for the Two Factors

	<i>n</i>	Form I		Form II	
		<i>F</i> ₁	<i>F</i> ₂	<i>F</i> ₁	<i>F</i> ₂
A&S men	326	.78	.72	.78	.76
A&S women	140	.74	.78	.72	.83
Engineers	181	.83	.82	.60	.69

Factor Analysis of Adjective Pairs

In addition to the major problem of learning the factor structure of pictures, it also was necessary to learn the factor structure of the polar adjectives used to rate each picture. Because for each picture we took the liberty of summing ratings over 15 pairs of polar adjectives, we implicitly assumed that such ratings are dominated by one general factor. To see if this was so, a factor analysis was performed on the 15 adjective pairs. First, mean scores over Ss were obtained for each of the 12 pictures on each of the 15 scales. Next, the 15 scales were intercorrelated (product-moment). Unities were inserted in the diagonal spaces, and centroid analysis was performed. The results of this analysis are presented in Table 3. Sixty per cent of the variance was accounted for by the first factor. The second accounted for 7 per cent and the third for 5 per cent. In other words, the first factor accounted for almost nine times as much variance as the second largest factor. Loadings on the first factor ranged from .68 to .84. We took this as evidence that our collection of polar adjectives is dominated by one general factor. The factor probably is closely akin to the factor of evaluation discussed by Osgood (1957) and his colleagues.

The Validity of the FIRS

Four studies have been completed which contribute evidence on the validity of the FIRS. The first study compared the self and peer ratings of groups scoring high and low on FIRS I and II. The criterion groups were the 40 highest-scoring and 40 lowest-scoring Ss (total scores) of 326 A&S males. These students were brought into the laboratory for further testing on an individual basis. The S was asked to rate himself, his closest personal friend, and the person he liked least, using the FIRS polar adjective scales as a rating

TABLE 3
Centroid Factor Analysis of the Polar Adjective Scales of
Form I for the 326 Arts and Science Men

% Variance	Factor No.	Polar Adjective Scale														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
59.6	1	.81	.78	.77	.76	.68	.82	.74	.76	.75	.74	.76	.78	.84	.82	.77
7.2	2	-.16	.32	-.32	.27	-.29	-.16	.40	-.25	.38	.20	.14	-.32	-.23	-.19	.21
4.8	3	.14	.29	.20	.33	.20	.22	.22	.20	-.13	-.32	-.24	-.15	-.16	.22	-.14

form. The means of the two groups on these three variables were compared by t tests. The highest-scoring group on the FIRS gave significantly higher ratings to themselves, their best friends, and the person whom they liked least ($p < .05$).

In the second study (Nelson and Izard, 1962), the FIRS I was administered to 42 female students preparing for church-related careers. Each person rated herself, the person in her group most like herself, and the person in her group least like herself. Correlations were computed between each of these variables and FIRS I+ and FIRS I-. FIRS I+ was correlated positively and significantly with the self-rating ($r = .47, p < .01$) and with the rating of peer most like self ($r = .40, p < .01$). FIRS I+ was significantly and negatively correlated with the rating of peer least like self ($r = -.32, p < .05$). None of the correlations for FIRS I- were significant.

In the third study (Izard and Lawton, 1961) FIRS II was administered to all the sophomore, junior, and senior members of a college fraternity. The fraternity was subdivided into six groups on the basis of year in college and their school affiliation (A&S, engineering). It was felt that in this way we were testing in "natural" groups where members could be expected to give reliable and accurate ratings of each other. After rating the pictures of FIRS II each S was asked to rate himself and every member of his group on the 15 polar adjective scales. FIRS II+ correlated significantly with self-rating ($r = .68, p < .01$) and with the S 's mean rating of his sub-group ($r = .68, p < .01$). The corresponding correlations for FIRS II- were significant but substantially smaller (.40, .44).

As part of a study concerned with the differential effects of positive and negative treatment on picture perception responses of normal and deviant S s, Izard and Jennings (1963) administered the FIRS I to 30 paratroopers who had received high personality-performance ratings from company officers and to 30 paratroopers who were in the stockade for at least their second offense. The FIRS I+ and FIRS I- scores of these criterion groups were compared by t test. The t for FIRS I+ was not significant. However, the FIRS I- successfully differentiated the two groups; t was 2.91, $p < .01$. For both FIRS I+ and FIRS I- the difference was in the expected direction—prisoners with two or more offenses had the lower mean.

Discussion

The factor analysis of the pictures gave convincing evidence that there are two relatively independent, reliable factors. One factor was based on responses to pictures that were judged to be positive affective stimuli, the other on pictures judged as negative affective stimuli. This result confirmed the expectation, based on the earlier studies with facial photographs, that more than one factor was operating. The finding of a "positive" and "negative" factor in the responses to pictures of human faces is quite consistent with our studies involving experimental manipulation of Ss' affect. In such studies we have shown that the quality of affect expressed by *E* made a substantial difference in perception and performance. Now, we have evidence that facial photographs rated as positive and negative affective stimuli are capable of evoking different traits or response tendencies.

As would be expected, the two factors correlated differently with other variables. The positive factor yielded higher correlations with ratings of self and peers "like self." The negative factor discriminated more successfully between adjusted and maladjusted groups. This suggested that the negative factor might tap responses less subject to conscious control.

The major import of these investigations has been to show that, when discussing individual differences in attitudes and feelings toward the "generalized-other," it makes a difference whether the "other" is a pleasant or unpleasant person. It is hoped that the procedure we have demonstrated for measuring these two traits will prove useful in future investigations of person-perception, particularly those concerned with the evaluative or affective component.

Summary

The assumption underlying this research was that evaluative polar adjective ratings of photographs of human faces would yield a measure of a basic personality dimension that has its roots in the individual's characteristic perceptual-affective response to people.

Judges were used to select 12 pictures that arouse positive feelings and 12 that arouse negative feelings. The pictures were divided into two sets (Forms I, II), each containing six positive and six negative pictures. Ss rated each picture, exposed for 90 seconds, on 15 nine-

point polar adjectives scales. The "First Impression Rating Scale" (FIRS) yields a total score (T), one based on the positive pictures (+), and one based on the negative pictures (-).

The Form I-Form II reliability of the FIRS T for 100 college students was .85. Test-retest reliability of FIRS I T for 34 females was .81.

Factor analyses were run on Forms I and II for 326 A&S men, 181 engineering students, and 140 A&S women. The factor analyses for the pictures indicated that there are two major factors, one relating to pictures of faces that manifest positive affect and one relating to pictures that manifest negative affect. The factor reliabilities and the interfactor correlations were satisfactory.

The factor analyses for the polar adjective rating scales revealed one general factor, the evaluative or affective dimension in person-perception.

Four validity studies have related FIRS scores to other perceptual and behavioral measures. FIRS+ correlated significantly with Ss' self-ratings and with the Ss' ratings of classmates and fraternity brothers. FIRS- differentiated between a group of paratroopers given high personality-performance ratings and a group in the stockade with two or more offenses on their records.

REFERENCES

- Beier, E. G., Izard, C. E., Smock, C. D., and Tougas, R. R. "Response to the Human Face as a Standard Stimulus." *Journal of Consulting Psychology*, XVII (1953), 126-131.
- Beier, E. G., Izard, C. E., Smock, C. D., and Tougas, R. R. "Response to the Human Face as a Standard Stimulus: A Re-Examination." *Journal of Consulting Psychology*, XXI (1957), 165-170.
- Izard, C. E. "Perceptual Responses of Paranoid Schizophrenic and Normal Subjects to Photographs of Human Faces." *American Psychologist*, VIII (1953), 372-373. (Abstract)
- Izard, C. E. "Paranoid Schizophrenic and Normal Subjects' Perceptions of Photographs of Human Faces." *Journal of Consulting Psychology*, XVII (1959), 119-124.
- Izard, C. E. and Jennings, J. R. "Personal Adjustment and Interpersonal Treatment as Variables in Stereoscopic Selective Perception." *American Psychologist*, XVIII (1963), 460.
- Izard, C. E. and Lawton, Mary. "The Interrelations of Perceptions of Self, Others, and Photographs of Human Faces." Unpublished study, 1961.
- Izard, C. E., Livsey, W. J., Cherry, E. S., Hall, G. F., Wall, Pat, and

- Bacon, Ruth. "Effects of Affective Picture Stimuli on the Learning, Perception, and Affective Scale Values of Previously Neutral Words." *ONR Technical Report No. 19*, Vanderbilt University, 1963.
- Nelson, Waudine and Izard, C. E. "Interpersonal Positive Affect in Women Students Who Have Selected Service Oriented Careers." *ONR Technical Report No. 7*, Vanderbilt University, 1962.
- Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning*. Urbana, Illinois: University of Illinois Press, 1957.
- Wehmer, G. and Izard, C. E. "The Effect of Self-Esteem and Induced Affect on Interpersonal Perception and Intellectual Functioning." *ONR Technical Report No. 10*, Vanderbilt University, 1962.

A VALIDATION OF HOWARD'S TEST OF CHANGE-SEEKING BEHAVIOR

GEORGE DOMINO¹

University of California, Berkeley

HOWARD, in 1961, began the development of a practical method for quantifying change-seeking behavior as a means for operationally testing the theoretical statements of White (1959), who formulated a theory of the need for competence, and of Fiske and Maddi (1961), who presented various hypotheses regarding the seeking of varied experience.

Although there has been an increasing number of experimental efforts concerning change-seeking behavior in humans, Howard's work represents one of the first comprehensive empirical efforts to operationally establish the existence and relevance of such a need.

Howard designed a number of visual paper-pencil mazes, each maze having several alternate paths to the goal or goals, all direct paths being equidistant from a goal, and all paths leading to a goal. These mazes are usually administered as a "filler" in a simple memory task, each *S* being asked to draw a line from the starting point to any goal for each of five identical mazes. The obtained score reflects the number of different path-segments taken.

A preliminary report (Howard, 1961) discussed the usefulness of these mazes in discriminating between psychiatric patients and tubercular and neurological patients. The psychiatric patients

¹ The writer, now at Fresno State College, is indebted to Dr. P. McReynolds for bringing Howard's work to his attention; to H. Diefenhaus for providing scoring instructions; to Miss D. Farmer for clerical assistance; to Prof. H. G. Gough for help and assistance in many phases of this project; to Dr. W. S. Gehman for his many helpful comments.

earned lower change-seeking behavior scores than did patients in the other two groups under study.

The present study was designed to obtain further data on the reliability and validity of one of these mazes, the Pyramid maze (PM) illustrated in figure 1. In particular, the study was designed

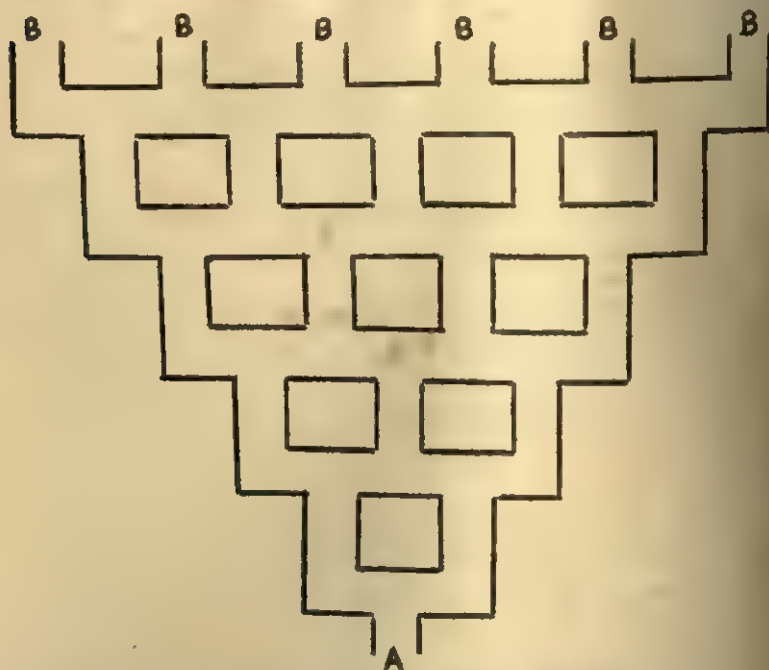


Figure 1. The Pyramid Maze (PM), indicating the starting point (A) and alternate goals (B).

to analyze the relationship between change-seeking scores as measured by the PM and various intellectual ability scores.

Procedure and Subjects

Procedurally, this research involved the administration of a variety of tests to 78 male freshmen enrolled in the author's introductory psychology courses at the University of San Francisco.² The tests used included the Army Beta, the Cattell Culture Fair, the College Vocabulary test (Gough & Sampson, 1954), the D-48 test (Gough & Domino, 1963), and the Gottschaldt Figures test (Crutch-

² Permission for this research, as well as research funds, were generously given by Dr. R. P. Vaughan, S.J., Director of Psychological Services, University of San Francisco.

field, 1952). These last three tests are not widely known, but were included as they seem quite promising.³

The College Vocabulary test is a 75-item multiple choice test. Two forms are available; form B was used in this study.

The D-48 test is a non-verbal test of intellectual ability. Although relatively unknown in the United States, it is widely used in Europe.

The Gottschaldt Figures test is a 20-item test of analytic perception involving the rapid identification of geometrical figures embedded in more complex gestalts. The form used in this study is the one developed by Crutchfield (1952).

In addition to these tests, Scholastic Aptitude Test scores were available for 62 Ss. These Ss had also participated, three months previously, in a testing program routinely administered to entering freshmen. Test results were thus available on the California Test of Mental Maturity (Sullivan, 1957) and the Ohio State University Psychological Test (Toops, 1940).⁴ Since not all Ss took all tests, the *N* varies from 62 to 78 in the analysis to be reported.

The PM was administered as a filler in a class demonstration. As can be seen in figure 1, this maze consists of a number of paths leading from a common starting point (A) to several alternate goals (B). Each of the paths is equally "correct" and equally distant from a goal, unless the *S* begins to circle back. Each direct path contains ten linear segments and a number of choice points. Change is defined as the number of segments traversed on a particular maze that differ from segments traversed on the maze immediately preceding.

Each *S* received a maze with the instructions to draw a line from the starting point to any goal. The maze was then collected and the entire procedure repeated four more times, each *S* traversing a total of five identical mazes. Change-seeking on the PM is operationally defined as the total of the four change scores (number of different segments traversed from maze to maze) times the number of mazes on which there is any change. The total score can thus vary from 0.00 indicating that the same path was consistently chosen on all

³ The D-48 and Gottschaldt Figures Test are published by the Consulting Psychologist Press, Palo Alto. The College Vocabulary Test is available from the authors, Institute of Personality Assessment and Research, University of California, Berkeley.

⁴ Dr. Robert Milligan, Director of the Testing and Counseling Center, University of San Francisco, very generously provided the results of this freshman testing program.

five mazes, to 16.00 indicating that each maze was traversed by a completely different path.

On the class meeting after the administration of the PM, a questionnaire was administered to the Ss. This questionnaire listed 20 extra-curricular events of general interest which had taken place during the previous three months. Each student was asked to check the number of events he had attended. It was hypothesized that responses to this questionnaire would provide a rough index of actual change-seeking behavior.

The 20 events listed were of a heterogeneous nature and, in order to avoid vitiating factors, were selected on the basis of these rational criteria: (1) none of the events charged admission, (2) all were held within walking distance of campus, (3) none was of a specialized nature, (4) none of the events was held during regular class hours, (5) not more than three events of the same nature (e.g. basketball games) were included, (6) attendance at these events was voluntary, and finally (7) none of the events was of such a social nature as to require a date (e.g. school dances were excluded).

In order to assess the reliability of the PM, 50 Ss were administered the PM twice: 27 Ss after a week's time, 23 Ss after approximately six months' time. Finally, at the end of the academic year, the grade point average for each S was obtained.

Results

The test-retest reliability of the PM was assessed by using a one-week and a six-month interval; product-moment coefficients of $+ .76$ and $+ .71$ were obtained. The split-half reliability was computed by comparing change scores on even versus odd trials on the first administration of the PM; the Spearman-Brown corrected reliability coefficient was $+ .91$. Reliability of the PM, whether of the test-retest or internal consistency variety, appears adequate.

The product-moment correlation coefficients between the PM and the other measures considered are presented in Table 1. Only two of these thirteen measures show a statistically significant relationship to the PM: the Verbal score on the SAT, $-.31$ ($p < .05$), and the questionnaire count of activities attended, $+ .42$ ($p < .01$).

Discussion

From the findings in Table 1, it seems there is little relation be-

TABLE 1

Correlations between the Variables Indicated and the Pyramid Maze

	N	r
Army Beta	62	.27
Cattell Culture Fair	68	.16
Calif. Test of Mental Maturity		
Language IQ	62	-.16
Non Language IQ	62	.01
Total IQ	62	-.09
College Vocabulary Test	72	.00
D-48	72	.13
Gottschaldt Figures Test	66	-.01
Ohio State Univ. Psychological Test	62	.00
Scholastic Aptitude Test		
Verbal score	62	-.31*
Mathematical score	62	-.17
Grade point average	78	-.02
Social Activities Questionnaire	77	.42**

* Significant at the .05 level.

** Significant at the .01 level.

tween the Ss' change-seeking scores on the PM and intellectual ability scores. This generalization would appear to apply equally well to non-verbal tests (the Army Beta, the Cattell, the D-48), to a direct measure of vocabulary (the College Vocabulary test), to a measure of analytic perception (Gottschaldt), and to standard tests of ability involving verbal and quantitative aspects (CTMM). Only one score, SAT-verbal, correlated significantly ($-.31, p < .05$) with the change-seeking score on the PM.

Although the absence of correlation with intellectual measures does not in itself assure the validity of the PM, it may be viewed as important in the presence of additional evidence. This evidence is found in the significant correlation coefficient obtained between change-seeking scores on the PM and number of extra-curricular activities in which Ss reported participation over a three months period, $+ .42, p < .01$. Thus "variety" of approach on the PM is associated with greater "variety" in one's social behavior.

It might be argued that the lack of a statistically significant relation between scores on the PM and measures of intellectual abilities reflects the homogeneity of the sample. However, the dispersion of scores seems broad enough to rule this out; for example, the SD on the Army Beta was 7.59 with IQ's ranging from 93 to 126; the SD on the CTMM was 11.94 with total IQ's from 77 to 148.

In summary, the following conclusions seem to be warranted:

(1) the PM test of Howard has adequate reliability, whether of the test-retest or internal consistency variety; (2) "variety" of approach on the PM is associated with greater "variety" in one's social behavior, while being relatively free of purely intellectual influence.

REFERENCES

- Crutchfield, R. S. *Gottschaldt Figures Test*. Berkeley, California: Institute of Personality Assessment and Research, University of California, Berkeley, 1952.
- Fiske, D. W. and Maddi, S. *Functions of Varied Experience*. Homewood, Illinois: Dorsey, 1961.
- Gough, H. G. and Domino, G. "The D-48 Test as a Measure of General Ability among Grade School Children." *Journal of Consulting Psychology*, XXVII (1963), 344-349.
- Gough, H. G. and Sampson, H. *The College Vocabulary Test*. Berkeley, California: Institute of Personality Assessment and Research, University of California, Berkeley, 1954.
- Howard, K. I. "A Test of Stimulus-Seeking Behavior." *Perceptual and Motor Skills*, XIII (1961), 416.
- Sullivan, Elizabeth T., Clark, W. W., and Tiegs, E. W. *California Test of Mental Maturity*, 1957 Edition, Monterey, California: California Testing Bureau, 1957.
- Toops, H. A. *Ohio State University Psychological Test*. Columbus, Ohio: Ohio College Association, 1919-1958.
- White, R. W. "Motivation Reconsidered: The Concept of Competence." *Psychological Review*, LXVI (1959), 297-333.

VALIDITY STUDIES SECTION

WILLIAM B. MICHAEL, Editor

University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor

Centinela Valley Union High School District
and the

University of Southern California

<i>Prediction of College Performance with the Myers-Briggs Type Indicator.</i> LAWRENCE J. STRICKER, HAROLD SCHIFFMAN, AND JOHN ROSS	1081
<i>Nonadditive Effects in the Prediction of Academic Achievement.</i> EDMOND MARKS AND JOSEPH E. MURRAY	1097
<i>Departmental Differentials in the Predictive Validity of the Graduate Record Examination Aptitude Tests.</i> GEORGE F. MADAUS AND JOHN J. WALSH	1105
<i>Personality Rigidity of Students Showing Consistent Discrepancies Between Instructor Grades and Term-End Examination Grades.</i> RODNEY T. HARTNETT AND CLIFFORD T. STEWART	1111
<i>Aptitude, Personality, and Achievement in Six College Curricula.</i> NORMAN M. CHANSKY	1117
<i>Validities and Intercorrelations of MMPI Subscales Predictive of College Achievement.</i> PHILIP HIMELSTEIN	1125
<i>Concurrent Validity of the Test of English as a Foreign Language for a Group of Foreign Students at an American University.</i> HENRY DIZNEY	1129
<i>Selection Techniques for Pakistani Postgraduate Students of Business.</i> GLEN GRIMSLEY AND GEORGE W. SUMMERS	1133
<i>The Predictive Validity of Eleven Tests at One State College.</i> RICHARD W. BOYCE AND R. C. PAXSON	1143
<i>The Predictive Validity of the School College Ability Test (SCAT) and the American College Test (ACT) at a Liberal Arts College for Women.</i> PAUL A. DE SENA AND LOUISE ANN WEBER	1149

<i>Validation of the Kahn Intelligence Tests.</i> ERNEST D. MCDANIEL AND WILLIAM T. CARSE	1152
<i>Uses of Cognitive and Non-Cognitive Test Measures in Sixty-four Private Liberal Arts Colleges: Implications for Predictive Validity and Assessment of Change.</i> ERNEST L. BOYER AND WILLIAM B. MICHAEL	1157

ANNOUNCEMENT REGARDING VALIDITY STUDIES

The VALIDITY STUDIES SECTION is published *twice a year*, once in the Summer issue and again in the Winter issue, for which the closing dates for receiving manuscripts are February first and August first, respectively. Although articles between two and eight printed pages are usually preferred, an occasional exception is made to publish articles of somewhat greater length.

Considerable flexibility exists concerning format as can be seen from a study of recently published articles. However, the model presented in the Spring, 1953, issue of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT still represents a close approximation to what is customarily published. The prospective contributor is encouraged to read the original announcement.

In order that the usual number of articles of other types may not be reduced, it is necessary to enlarge the journal and to charge the authors for most of the publishing costs. For a running page of printed text the cost is fifteen dollars per page with extra charges for tables and complex material. Each author receives 100 free reprints.

Manuscripts should be sent to:

Dr. William B. Michael
Professor of Education and Psychology
University of California, Santa Barbara
Santa Barbara, California

PREDICTION OF COLLEGE PERFORMANCE WITH THE MYERS-BRIGGS TYPE INDICATOR¹

LAWRENCE J. STRICKER

Educational Testing Service

HAROLD SCHIFFMAN

Duke University

AND

JOHN ROSS

University of Western Australia

THE aim of this article is to assess the ability of the Myers-Briggs Type Indicator (Myers, 1962b) to predict two aspects of college performance—freshman year grade-point average (*GPA*) and dropout—through the use of the Indicator's continuous scores as well as measures intended to reflect its interdependent, dichotomous type categories, and correcting the obtained correlations, where possible, for the multivariate selection inherent in most samples of college students.

The Myers-Briggs Type Indicator is a self-report inventory which is intended to measure variables stemming from the Jungian personality typology. It consists of four scales: Extraversion-Introversion (*E-I*), Sensation-Intuition (*S-N*), Thinking-Feeling (*T-F*), and Judging-Perceiving (*J-P*). The *E-I* scale is

¹ Thanks are due Dr. C. Hess Haagen of Wesleyan University and Dr. John R. Weir of California Institute of Technology for furnishing the raw data used in the studies reported in this article, and Mr. Roald Buhler of Princeton University for writing a computer program to carry out the corrections of correlation matrices for multivariate selection.

A table reporting the scores used in the four-scale contingency predictions has been deposited with the American Documentation Institute. Order Document No. 8539, remitting \$1.25 for 35-mm. microfilm or \$1.25 for 6 x 8 in. photographs.

presumed to measure interest in things and people or concepts and ideas; the *S-N* scale, tendencies to perceive through the usual sensory processes or indirectly, via the unconscious; the *T-F* scale, tendencies to judge (or evaluate) phenomena rationally and impersonally or subjectively and personally; and the *J-P* scale, tendencies to reach conclusions about phenomena or to become aware of them.

These scales were expressly developed to classify people into type categories (e.g., classification as an extravert, an introvert, or, in those cases where the two tendencies are equal, "indeterminate") which would have real meaning. The cutting (or "zero") points used in making these classifications were so chosen that those people who are on one side of a scale's cutting point, and, hence, in one type category, are presumed to be qualitatively different from those who are on the other side of it, and hence, in the opposite type category. In addition to these categorical classifications, continuous scores for each scale can be derived by arbitrarily considering one end of the scale high (Stricker and Ross, 1963, p. 287).²

Previous studies by the present authors (Ross, 1961; Ross, 1963; Stricker and Ross, 1963; Stricker and Ross, 1964a; Stricker and Ross, 1964b), as well as others (Howarth, 1962; Lord, 1958; Myers, 1962a; Myers, 1962b; Saunders, 1960), have assessed various aspects of the construct validity of the Indicator. Some data also exist concerning the Indicator's utility in predicting school performance and other socially important variables. Most studies relevant to school performance have reported the correlations of each of the Indicator's scales with *GPA* (Myers, 1962b; Nichols and Holland, 1963; Stricker and Ross, 1964a). The median correlations over 15 high school and college samples (Myers, 1962b) ranged from .06 for the *T-F* scale to .12 for the *S-N* scale. The scales' correlations with stu-

² After the studies reported in this article were completed, the Indicator's manual appeared, changing the scoring system so as to eliminate indeterminate type categories. This goal was accomplished by combining a scale's indeterminate type category with one of the two other type categories on that scale. The original continuous scores were also linearly transformed. The use of the new scoring would have no appreciable effect on these studies' contingency prediction results, in view of the small number of *Ss* in the indeterminate type categories, and would not alter the correlations reported for the continuous scores in the present studies, although it does affect the means and standard deviations.

dents' own reports of their extra-curricular achievement have also been reported (Nichols and Holland, 1963). The *E-I* scale correlated significantly with achievement in leadership and dramatic activities; several scales, consistently including *T-F*, correlated significantly with literary, musical, and graphic art achievement; and none correlated with scientific achievement. In addition, analysis of variance has been used to investigate mean differences in *GPA* and over-under achievement between the two major type categories on each scale, and interactions between the four Indicator scales with respect to these two criteria (Stricker and Ross, 1964a). No significant effects were found. There are no published data about the multiple correlations of the four Indicator scales with these criteria, or about the Indicator's ability to predict dropout.

With the exception of the analysis of variance findings, these previous studies have employed the Indicator's continuous scores, and have analyzed the four scores separately, although the Indicator's underlying theory postulates interacting, dichotomous type categories. A more appropriate test of this inventory's validity might use predictions based on interdependent type classifications, since they would be more nearly consistent with the intent of both the Indicator and the theory underlying it.

The standard statistics that measure the relationship between categorical variables and continuous variables cannot be readily used to make predictions or to measure the joint effects of several variables. However, a contingency-table procedure described by Lykken and Rose (1963) generates scores from categorical variables that can be used in standard correlation and regression routines, and is applicable to any number of categorical variables, whether stemming from one basis of classification (e.g., extraversion-introversion) or several joint ones (e.g., extraversion-introversion, sensation-intuition, and sex). This procedure simply involves assigning as a score for each subject in a particular category the mean criterion score achieved by subjects in that category. To avoid capitalizing on chance, the mean score should be established in an independent sample. As applied to the Indicator, this procedure requires the classification of subjects into the 81 possible categories that are produced by considering all combinations of the three type categories on the four scales. The results obtained in this way should reflect the interdependence of the Indicator scales—in-

teractions and linear combination effects—and should be consistent with the intent of the Indicator and the typological theory. This procedure can also be used separately with each scale, by classifying subjects into the three type categories on the scale, and the effects of each scale can be appraised.

Prediction studies typically employ samples that have been selected in some explicit way, and the previous studies of this kind with the Indicator have been based on such samples. As Thorndike (1947) demonstrated, this selection may drastically affect the relationships among the variables used in selection, as well as any other variables that are related to those used in selection. One common result is that new predictors may appear to be more valid than existing ones, merely because the existing predictors were used in selection. Corrections for multivariate selection need to be applied to the prediction data for the Indicator if its validity is to be precisely estimated, both in absolute terms and in relation to the ability of existing predictors.

Method

Samples and Bases of Selection

a. Wesleyan—A group of 225 men from the 254-man freshman class entering Wesleyan University in 1959. (The other men were excluded because scores on variables in a related study were not available for them.)

b. California Institute of Technology—The entire 201-man freshman class entering California Institute of Technology (Caltech) in 1958, and the 1616 men, including the 201, who applied for admission to that class. The 1415 who were not in the entering class either were rejected or were accepted but did not register. The half of the total applicant group with the lowest scores on a composite of five predictors were initially rejected in the selection process, without an interview. The five predictors were the *Scholastic Aptitude Test's* verbal (SAT-V) and mathematical (SAT-M) subtests (College Entrance Examination Board, 1964; Dyer and King, 1955) and the College Board achievement tests in *Physics*, *Advanced Mathematics*, and *Chemistry* (College Entrance Examination Board, 1963; Dyer and King, 1955). Decisions to accept or reject the other applicants were made after an interview, in which inter-

viewers had access to the composite score as well as the scores for the tests used in the composite, the College Board achievement test in *English*, and other pertinent information.

Criteria

The criteria were freshman-year *GPA*³ and dropout (available for Caltech only)—0 = staying, 1 = leaving.

Predictors

Existing Predictors. *SAT-V* and *SAT-M* scores, as well as high school rank⁴, were obtained from school records for the two entering classes. The *SAT* scores, but not high school rank, were also available for the 1415 men not entering Caltech. Available scores on the College Board achievement tests were also obtained for the two Caltech groups. A complete set of *SAT* and College Board achievement test scores was not available for all men because applicants were not required to take all tests.

Indicator Measures. The Indicator had been administered to the two entering classes at the beginning of the school year. The following kinds of scores were available for both:

a. Continuous Scores—A score for each scale was obtained by arbitrarily considering one end of the scale high. In this study, the *I*, *N*, *F*, and *P* ends of the scale are high and the *E*, *S*, *T*, and *J* ends are low. Hence, high scores on the *E-I*, *S-N*, *T-F*, and *J-P* scales signify, respectively, tendencies towards introversion, intuition, feeling, and perceiving; low scores on the scales signify extraversion, sensation, thinking, and judging.⁵

³ Both schools assign letter grades, but quantify them differently. At Wesleyan, A = 95, B = 85, C = 75, D = 65, and F = 45; at Caltech, A = 4, B = 3, C = 2, D = 1, and F = 0. In the present study, each school's own grading system was retained, but note that the two are linearly related over all but the lowest interval of each grading scale.

⁴ Students' class standings in high school (e.g., third in a class of 200) were transformed into standard scores (on the assumption of a normal distribution) with a mean of 13 and a standard deviation of 4, with high scores signifying high performance. The few students in each class for whom these data were not available were assigned the median high school rank of their entering class.

⁵ Note that in the present study, as well as in a previous one (Stricker and Ross, 1964b), the convention used for obtaining continuous scores described in the Indicator's manual (Myers, 1962b) was followed. In two earlier articles by the present authors (Stricker and Ross, 1963; Stricker and Ross, 1964a), the opposite scoring convention was followed: the *E*, *S*, *T*, and *J* ends of the scales were high.

b. Four-scale *GPA* Contingency Prediction—In each entering class, each student in each of the 81 possible four-variable type categories was assigned the mean freshman *GPA* of all the students in the same category.

An additional score, the four-scale dropout contingency prediction—identical to the *GPA* contingency prediction, but based on the proportion leaving during freshman year—was available only for the Caltech entering class.

Procedures

The predictors and criteria were intercorrelated separately for the two entering classes. Product-moment correlations were calculated throughout, with the exception that point-biserial correlations were used with the dichotomous Caltech dropout variable.⁶ The intercorrelations of the predictors and criteria for the total Caltech group were estimated from their intercorrelations in the entering class by (a) using the *SAT* and achievement test scores common to the two Caltech groups, and (b) following the procedures for correcting for multivariate selection described by Gulliksen (1950). Since scores on all the *SAT* and College Board achievement tests were not available for all students in the two Caltech groups, the *Ns* on which these correlations are based vary. In the entering class, the median *N* for the correlations between the *SAT* and College Board achievement tests was 146, with a range of 44 to 201. In the total group, the median was 1097, with a range of 286 to 1587.

Within the three samples, multiple correlations (*Rs*) of various sets of predictors with each criterion were computed and corrected for shrinkage (Wherry, 1931). The gain in validity resulting from the addition of the Indicator scores to the common predictors—

⁶ In view of the possibility that the regressions underlying these correlations were nonlinear, the following scatterplots within each entering class were inspected:

- (a) each predictor (except the College Board achievement tests) against the appropriate criteria;
- (b) each Indicator scale against each of the others;
- (c) *SAT-V*, *SAT-M*, and high school rank against each other;
- (d) each Indicator measure against *SAT-V*, *SAT-M*, and high school rank.

Seven of the 150 regressions seemed to be nonlinear, and one of these involved the regression of a criterion on a predictor, but in this instance the nonlinearity did not seem to be marked.

SAT-V, *SAT-M* and high school rank—was assessed by *F* tests of differences between the *R*s for the common predictors and the *R*s for the common predictors plus the various kinds of Indicator scores.

The correlations of the four-scale contingency predictions with their corresponding criteria may be inflated to some degree because the contingency predictions are based on the actual means or proportions for the same groups rather than for independent samples. Hence, these correlations should be considered as upper limits for those that would be obtained by cross-validation studies in which the present scores are used to predict the same criteria in independent, but similar, samples.

It should also be noted that the estimated correlations for the total Caltech group may be in error to some unknown extent because the actual selection was not based entirely on the five *SAT* and College Board tests used in the estimation procedure. Although these variables were the only basis of selection in the half of the applicant group who were rejected because their composite scores on these tests placed them in the lower half of the group, additional variables may have been operative in the selection by the school and in the self-selection by the applicants within the half of the group with the higher composite scores. The amount of error introduced by these extraneous variables seems limited, for, at most, only half the sample could be affected. Moreover, it is likely that the tests used in the estimation procedure influenced, at least in part, the selection decisions within that subsample.

TABLE 1

Intercorrelations of Predictors and Criteria for the Wesleyan Class (N = 225)

Variable	Mean	S.D.	2	3	4	5	6	7	8	9
1. <i>SAT-V</i>	628.18	77.31	38	27	27	34	-10	08	17	33
2. <i>SAT-M</i>	656.04	76.09		39	20	22	-15	-03	21	24
3. H. S. Rank	18.20	3.35			13	11	-12	-10	30	50
4. <i>E-I</i> Scale	2.38(E)	13.49				00	-08	02	40	18
5. <i>S-N</i> Scale	4.76(N)	13.17					14	34	01	07
6. <i>T-F</i> Scale	.00(X)	10.46						24	-17	01
7. <i>J-P</i> Scale	1.72(P)	12.96							-41	-24
8. Four-Scale GPA										48
Contgcy	81.03	3.02								
9. GPA	81.02	6.25								

Note—Decimal places for correlation coefficients have been omitted.

TABLE 2
Intercorrelations of Predictors and Criteria for the Caltech Class (N = 201) and Total Group (N = 1616)

Variable	Class		Total Group																
	Mean	S.D.	Mean	S.D.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. SAT-V	668.21	64.72	594.21	93.68		05	27	-05	-02	59	10	11	20	-12	17	00	-02	08	05
2. SAT-M	753.89	39.70	668.15	83.94	49		18	35	-01	11	17	03	07	-05	-02	20	15	31	-20
3. Physics Ach. Test	693.22	47.59	609.63	83.75	59	59		22	18	21	05	06	08	-07	-03	05	11	24	-13
4. Mathematics Ach. Test	765.48	38.80	664.43	89.80	46	71	61		07	17	05	03	13	-03	-05	14	07	26	-20
5. Chemistry Ach. Test	724.44	55.94	617.30	99.56	58	59	77	64		11	-02	-09	08	-14	12	-11	05	13	05
6. English Ach. Test	647.74	55.39	585.61	84.59	72	51	50	48	50		16	09	18	-12	07	16	06	26	-06
7. H. S. Rank	21.22	2.60	—	2.76	25	35	21	22	20	33		-05	06	-10	-10	12	-01	29	-10
8. E-I Scale	3.94(I)	13.33	—	13.34	11	06	04	06	00	11	-03		-13	-15	-12	20	11	07	02
9. S-N Scale	13.77(N)	8.63	—	9.99	46	41	41	47	45	38	18	-09		00	24	21	09	10	-07
10. T-F Scale	5.81(T)	9.94	—	10.73	-35	-31	-34	-29	-38	-30	-19	-16	-18		23	-11	-11	-07	09
11. J-P Scale	1.97(J)	14.39	—	14.76	29	12	15	12	22	16	-04	-11	30	13		-34	-26	-13	09
12. Four-Scale GPA Contingy	2.74	.23	—	.25	12	34	14	25	10	32	21	20	27	-18	-28		53	37	-19
13. Four-Scale Dropout Contingy	.12	.11	—	.12	17	35	30	26	27	24	10	12	22	-21	-18	56		19	36
14. GPA	2.74	.64	—	.90	47	69	61	64	60	56	41	09	40	-30	04	44	36		-50
15. Dropout	.11	.32	—	.35	-16	-42	-31	-40	-26	-26	-20	00	-23	21	02	-28	44	-60	

Note.—The correlations for the entering class appear above the diagonal and those for the total group appear below. Decimal points for correlation coefficients have been omitted.

Results

The means, standard deviations, and intercorrelations of the predictors and criteria at Wesleyan appear in Table 1. These statistics for the entering class and for the total group at Caltech appear in Table 2.

Correlations between Indicator Measures

In the Wesleyan class, the R of the four Indicator scales with the four-scale contingency prediction of GPA was .60 ($p < .01$). In the Caltech entering class, the R s were .47 ($p < .01$) with the GPA contingency prediction, and .29 ($p < .01$) with the dropout contingency prediction. In the total Caltech group, the R s were .51 ($p < .01$) and .37 ($p < .01$).

Correlations with Criteria

Wesleyan GPA. The Indicator's $E-I$ and $J-P$ scales correlated significantly with GPA . Their correlations were .18 ($p < .01$) and $-.24$ ($p < .01$), respectively.

The four-scale contingency prediction of GPA correlated .48 ($p < .01$) with the GPA criterion, larger than was the R of .33 ($p < .01$) between the four Indicator scales and this criterion. The R of $SAT-V$, $SAT-M$ and high scale rank with this criterion was .54 ($p < .01$).

Caltech GPA. The correlations of the various Indicator measures with the criteria were generally lower in the Caltech entering class than in the Wesleyan entering class, but the differences between the latter sample and the total group at Caltech were less pronounced.

None of the Indicator's scales correlated significantly ($p > .05$) with GPA in the entering class. In the total group, however, all the corresponding correlations were significant ($p < .05$), and, except for the $J-P$ scale, were larger than the correlations in the entering class. The correlations in the two groups were .07 and .09 for the $E-I$ scale, .10 and .40 for the $S-N$ scale, and $-.07$ and $-.30$ for the $T-F$ scale.

The correlations of the four-scale contingency prediction of GPA with the GPA criterion were .37 ($p < .01$) in the entering class and .44 ($p < .01$) in the total group, and the corresponding R s of the Indicator scales with GPA were .16 ($p > .05$) and .47 ($p < .01$).

The *Rs* of *SAT-V*, *SAT-M*, and high school rank with *GPA* were .39 ($p < .01$) and .73 ($p < .01$).

Caltech Dropout. None of the Indicator's scales correlated significantly ($p > .05$) with dropout in the entering class, but the *S-N* and *T-F* scales were significantly correlated in the total group. Their correlations of $-.23$ ($p < .01$) and $.21$ ($p < .01$) were also the two that were larger in the total group than in the entering class.

The correlations of the four-scale contingency prediction of dropout with the dropout criterion were .36 ($p < .01$) in the entering class and .44 ($p < .01$) in the total group, and were larger than the corresponding *Rs* of .08 ($p > .05$) and .29 ($p < .01$) between the Indicator scales and this criterion. The *Rs* of *SAT-V*, *SAT-M*, and high school rank with this criterion were .19 ($p > .05$) and .42 ($p < .01$).

Incremental Validity of Indicator

Table 3 reports the *Rs* of the various predictor combinations with the *GPA* criterion in the Wesleyan class and the *GPA* and dropout criteria in the two Caltech groups. These *Rs* were based on *SAT-V*, *SAT-M*, and high school rank only, as well as on these predictors plus each kind of Indicator measure.

Wesleyan. The *Rs* with *GPA* at Wesleyan that resulted when the

TABLE 3
*Multiple Correlations of Indicator Measures and Standard Predictors
with Criteria at Wesleyan and Caltech*

Predictor Combination	No. Pre- dictors	Wesleyan Class (<i>N</i> = 225)	Caltech Class (<i>N</i> = 201)		Caltech Total Group (<i>N</i> = 1616)	
		<i>GPA</i>	<i>GPA</i>	Drop- out	<i>GPA</i>	Drop- out
Indicator Scales	4	.33**	.16	.08	.47**	.29**
Four-Scale Contgcy Prediction	1	.48**	.37**	.36**	.44**	.44**
Indicator Scales, <i>SAT</i> Scales, and H.S. Rank	7	.59**	.39**	.19	.74**	.45**
Four-Scale Contgcy Prediction, <i>SAT</i> Scales, and H.S. Rank	4	.63**	.48**	.38**	.76**	.53**
<i>SAT</i> Scales and H.S. Rank	3	.54**	.39**	.19	.73**	.42**

** Significant at .01 level.

Indicator scales and the four-scale contingency prediction were each added separately to *SAT-V*, *SAT-M*, and high school rank were both significantly ($p < .01$) larger than was the R of .54 with this criterion for these standard predictors. The largest increase (.09) was produced by the contingency prediction.

Caltech. The order of magnitude of the R s was similar for the two Caltech criteria—*GPA* and dropout. For each criterion in the entering class, only the R produced by the addition of the four-scale contingency prediction to the standard predictors was significantly ($p < .01$) larger than was the R for the standard predictors only. The increase was .09 for *GPA* and .19 for dropout; the R s of the standard predictors with these criteria were .39 and .19, respectively.

For each criterion in the total group, however, both R s produced by the separate addition of the two Indicator measures to the standard predictors were significantly ($p < .01$) larger than was the corresponding R for the standard predictors only. The largest increase for each criterion was produced by the four-scale contingency prediction. This increase was .03 for *GPA* and .11 for dropout; the R s of the standard predictors with these criteria were .73 and .42, respectively.

Similar increases in R s in the two Caltech groups were obtained by the separate addition of the two kinds of Indicator measures to the three standard predictors and the four College Board achievement tests.

Other Correction for Selection Effects

Some of the differences between the correlations of the Indicator measures for the two Caltech groups have already been indicated. There were a number of other differences between the correlations for the two groups.

The correlations in the total group were generally higher than were those in the entering class. The largest differences occurred for the *SAT* and achievement tests. The median correlations (disregarding sign) between these measures were .17 in the entering class and .59 in the total group, and their median correlations with the criteria were .15 and .42. The differences were not so marked for the Indicator measures, but were largest for the *S-N* and *T-F* scales. The median correlations of these two scales with the criteria were

.08 in the entering class and .26 in the total group; the median correlations for the *E-I* and *J-P* scales were .08 and .03. In addition, the *S-N* and *T-F* scales had median correlations of .10 in the entering class and .38 in the total group with the *SAT* and achievement tests; the correlations for the other Indicator scales were .06 and .12. The median correlations of the four-scale contingency predictions with their own criteria were .36 in the entering class and .44 in the total group, and their median correlations with the *SAT* and achievement tests were .09 and .26.

Discussion

The Indicator's scales had some ability to predict the two criteria that were studied, even in the selected samples, and this ability varied sharply with the criterion and the nature of the sample.

These correlations between the Indicator's scales and *GPA* in the entering classes are similar to those reported for entering classes at other liberal arts colleges and engineering schools (Myers, 1962b; Stricker and Ross, 1964a), and National Merit Finalists in a cross-section of colleges (Nichols and Holland, 1963). The correlations for the *S-N* and *T-F* scales in the total Caltech group are higher than are the previously reported correlations, probably because of the effects of selection in the other groups. Furthermore, the Indicator's correlations with *GPA* in the entering classes are roughly similar to the correlations of other personality measures with this criterion which have been reviewed by Fishman and Pasanella (1960), Garrett (1949), Harris (1940), and Travers (1949). The median correlations were reported to be .09 by Garrett and .22 by Fishman and Pasanella.

While the Indicator's scales have some predictive validity, the crucial consideration from the pragmatic standpoint of prediction is the increment in validity produced by adding these Indicator measures to existing predictors. The improvement in prediction of the two criteria that does result from the addition of the Indicator scales to the standard prediction combination—the increase in the *R* is .05 at Wesleyan and a maximum of .03 in the total Caltech group, but there is no increase in the Caltech entering class—is slight, and is about the same as generally obtained in grade prediction studies when other personality scales are added to the usual intellectual predictors (Fishman and Pasanella, 1960).

The finding that the four-scale contingency predictions generally did have greater predictive validity than did the Indicator scales, either singly or in combination, and frequently had greater validity than the standard predictor combination, together with the finding that these contingency measures cannot be adequately predicted from the Indicator's scales suggest that this procedure may be promising. These results also raise the possibility that the Indicator scales do display significant interaction effects for some variables, contrary to the previous analysis of variance findings for *GPA* and for over-under achievement (Stricker and Ross, 1964a). Cross-validation of the contingency predictions in other Wesleyan and Caltech classes is needed to eliminate the inflation inherent in the present correlations for the contingency predictions.

Incidentally, at least in the two entering classes, the correlations of the Indicator's measures with over-under achievement (conventionally defined as actual *GPA* less *GPA* predicted from *SAT-V* and *SAT-M*) would be similar to their correlations with *GPA*, for the two criteria would be highly correlated.⁷

Another finding of some interest was the striking difference between the correlations for the entering Caltech class and those for the total group. Predictably, the largest differences occurred for the variables that had been explicitly used in selection—the *SAT* and achievement tests—and other sizable differences, though not so large as the previous ones, occurred for the implicit selection variables—the *S-N* and *T-F* scales—that were related to the explicit selection variables.

Summary

This study investigated the ability of the Myers-Briggs Type Indicator, a self-report inventory, to predict grades and dropout at Wesleyan and Caltech, both for classes and, using estimates, for

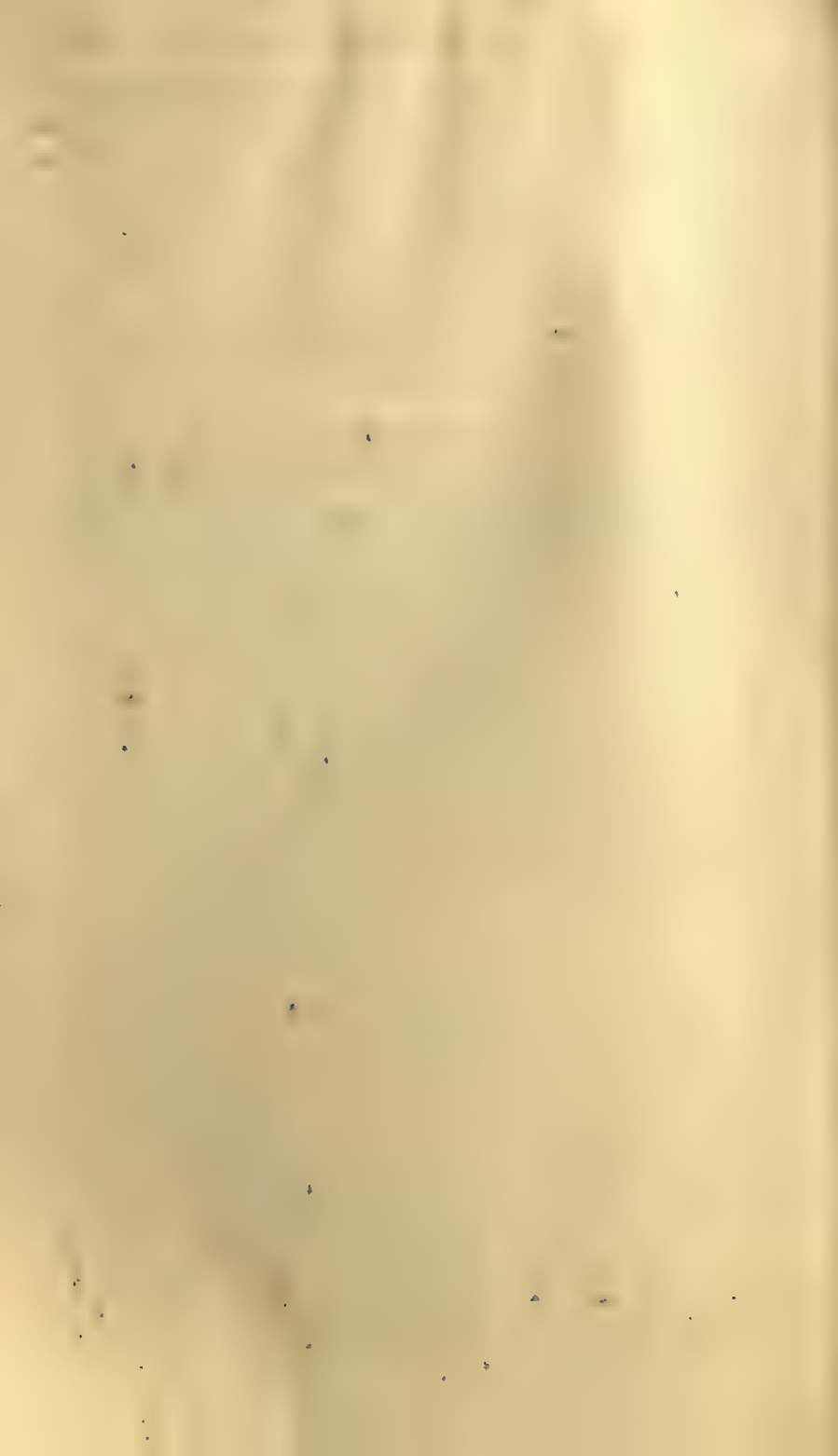
⁷ It is well known that r^2 can be interpreted as the proportion of variance shared by a pair of variables. The variance common to *GPA* and over-under achievement (defining the latter as actual *GPA* minus *GPA* predicted from the *SAT* scales) would be unity, except for the existence in *GPA* of variance associated with *SAT*. Hence, the variance shared by *GPA* and over-under achievement is $1 - R^2(\text{SAT})(\text{GPA})$ and the correlation between the two is $\sqrt{1 - R^2(\text{SAT})(\text{GPA})}$. The R s of *SAT-V* and *SAT-M* with *GPA* were .35 in the Wesleyan class, .32 in the Caltech entering class, and .71 in the total Caltech group, so the correlations between *GPA* and over-under achievement in the three groups would be .94, .95, and .70.

Caltech applicants. Continuous scores derived from the Indicator's four scales had some ability to predict the criteria, and this ability varied with the criterion and the sample. A contingency measure that reflects the Indicator's interdependent, dichotomous type categories generally had greater predictive validity than did the continuous scores, but its correlations may be inflated. Adding the Indicator's continuous scores to the SAT scales and high school rank produced a slight but significant improvement in prediction in the Wesleyan class and the Caltech applicant group, but not in the Caltech class.

REFERENCES

- College Entrance Examination Board. *A Description of the College Board Achievement Tests*. Princeton, N. J.: Author, 1963.
- College Entrance Examination Board. *A Description of the College Board Scholastic Aptitude Tests*. Princeton, N. J.: Author, 1964.
- Dyer, H. S. and King, R. G. *College Board Scores: Their Use and Interpretation*. Princeton, N. J.: College Entrance Examination Board, 1955.
- Fishman, J. A. and Pasanella, Ann K. "College Admission-Selection Studies." *Review of Educational Research*, XXX (1960), 298-310.
- Garrett, H. F. "A Review and Interpretation of Investigations of Factors Related to Scholastic Success in Colleges of Arts and Science and Teachers Colleges." *Journal of Experimental Education*, XVIII (1949), 91-138.
- Gulliksen, H. *Theory of Mental Tests*. New York: Wiley, 1950.
- Harris, D. "Factors Affecting College Grades: A Review of the Literature, 1930-1937." *Psychological Bulletin*, XXXVII (1940), 125-166.
- Howarth, E. "Extroversion and Dream Symbolism: An Empirical Study." *Psychological Reports*, X (1962), 211-214.
- Lord, F. M. "Multimodal Score Distributions on the Myers-Briggs Type Indicator—I." Research Memorandum 58-8. Princeton, N. J.: Educational Testing Service, 1958.
- Lykken, D. T. and Rose, R. "Psychological Prediction from Actuarial Tables." *Journal of Clinical Psychology*, XIX (1963), 139-151.
- Myers, Isabel B. "Inferences as to the Dichotomous Nature of Jung's Types, from the Shape of Regressions of Dependent Variables upon Myers-Briggs Type Indicator Scores." *American Psychologist*, XVII (1962), 364. (Abstract) (a)
- Myers, Isabel B. *Manual (1962), the Myers-Briggs Type Indicator*. Princeton, N. J.: Educational Testing Service, 1962. (b)
- Nichols, R. C. and Holland, J. L. "Prediction of the First Year College Performance of High Aptitude Students." *Psychological Monographs*, LXXVII (1963), No. 7 (Whole No. 570).

- Ross, J. "Progress Report on the College Student Characteristic Study: June 1961. Research Memorandum 61-11. Princeton, N. J.: Educational Testing Service, 1961.
- Ross, J. "The Relationship between the Myers-Briggs Type Indicator and Ability, Personality and Information Tests." Research Bulletin 63-8. Princeton, N. J.: Educational Testing Service, 1963.
- Saunders, D. R. "Empirical Evidence for a Rational Correspondence between the Personality Typologies of Spranger and of Jung." *American Psychologist*, XV (1960), 459. (Abstract)
- Stricker, L. J. and Ross, J. "Intercorrelations and Reliability of the Myers-Briggs Type Indicator Scales." *Psychological Reports*, XII (1963), 287-293.
- Stricker, L. J. and Ross, J. "An Assessment of Some Structural Properties of the Jungian Personality Typology." *Journal of Abnormal and Social Psychology*, LXVIII (1964), 62-71. (a)
- Stricker, L. J. and Ross, J. "Some Correlates of a Jungian Personality Inventory." *Psychological Reports*, XIV (1964), 623-643. (Monogr. Suppl., 1964, no. 6-V14). (b)
- Thorndike, R. L. (Editor) *Research Problems and Techniques. AAF Aviation Psychology Program Research Reports, Report No. 3.* Washington, D. C.: U. S. Government Printing Office, 1947.
- Travers, R. M. W. "Significant Research on the Prediction of Academic Success." In Wilma T. Donahue, C. H. Coombs, and R. M. W. Travers (Editors), *The Measurement of Student Adjustment and Achievement.* Ann Arbor, Michigan: University of Michigan Press, 1949, 147-190.
- Wherry, R. J. "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation." *Annals of Mathematical Statistics*, II (1931), 440-457.



NONADDITIVE EFFECTS IN THE PREDICTION OF ACADEMIC ACHIEVEMENT

EDMOND MARKS

The Pennsylvania State University

AND

JOSEPH E. MURRAY

United States Army

HIGH school average (*HSA*) has long been considered as that variable which singly provides the most information about future academic performance at the college level as measured by college grade point average (*GPA*) (Segel, 1934; Travers, 1949) or attrition (Lindsay, Marks, and Hamel, 1964). This predictive relationship has been attributed to the hypothesized multidimensional nature of *HSA*, which has been assumed to reflect such factors as scholastic ability, motivation, the development and utilization of skills relevant to scholastic performance, and others (Fishman, 1962). As such, *HSA* is treated as a vector, the several components of which reflect factors important to college academic achievement. Demonstration of the informational properties of *HSA* is typically undertaken by utilizing statistical models such as bivariate or *k*-variate regression where the equation is assumed linear in the parameters and in the independent variables.

Where two secondary schools differ in quality of students, quality of instruction, competitiveness, or grading standards, there is reason to question whether the same *HSA* reported by each can be considered equivalent. Many attempts have been made, ranging from the completely intuitive to the statistically sophisticated, to improve the predictor properties of *HSA* by taking into considera-

tion the variation in standards among secondary schools (Bloom and Peters, 1961; Chauncey and Frederiksen, 1951; Lindquist, 1963; McClelland, 1942).¹ Thus, the informational properties of *HSA* are increased by knowing certain characteristics of the secondary school. These attempts are, however, characterized by transformations on *HSA* which assume that there is no interaction between *HSA* and those variables used in the adjustment.

It occurred to the authors that the relationship between college academic achievement and a composite of *HSA* and secondary school characteristics was much more subtle than that suggested by a linear, additive approach. Whenever a given secondary school is highly oriented to preparing students for college work, one might expect the characteristics of that school—particularly the behaviors of the administration, faculty, and students—to be more similar to those of the typical college or university than are the characteristics of a secondary school not so oriented. Some secondary schools are more “like” colleges or universities than are others. This suggests that *HSA* does not uniformly provide the same amount of information about college academic performance for all subjects (*Ss*), but provides differential amounts as a function of the similarity between secondary school and college characteristics. This was the hypothesis under investigation in this study.

Method

Subjects

Data were collected on all baccalaureate degree students entering The Pennsylvania State University in the fall term 1963 for whom first-term grade point averages were available ($N = 3722$). Because of the possible effects of type of program upon the hypothesis under investigation, the *Ss* were split into two groups; (1) *Ss* whose college program required mathematics at least through integral calculus (Science, $N = 1483$), (2) *Ss* whose program did not have this requirement (Non-Science, $N = 2239$). The median age of the total group was 18, with most of the *Ss* being recent high school graduates.

¹ Unpublished data in the files of Student Affairs Research, The Pennsylvania State University, indicate that making a simple linear adjustment upon *HSA* in terms of quality of the secondary school improves the correlation between *HSA* and college academic achievement.

Procedure

For each individual *S*, measures of high school achievement, of first-term college achievement, and of a similarity index between the secondary school and the college were obtained. High school and college achievement were computed separately but in the same way by summing the product of the quality point ($A = 4$, $B = 3$, $C = 2$, $D = 1$, and $F = 0$) and the number of credits for a given course and dividing by the total number of credits. The percentage of a secondary school's graduating class enrolling in an accredited college or university was employed as a measure of high school-college similarity. Although high school-college similarity is assumed multidimensional, the percent going to college appears to represent an adequate composite of the various components of similarity (Laughlin, 1963).

Ss in both groups, i.e., Science and Non-Science, were rank ordered in terms of the percentage of the graduating class going to college, and the distribution of values was plotted. Based on this distribution, each group was split into four subgroups so as to insure approximate equality of the variability of the percentage going to college for the four subgroups. The score ranges established were: Group I, 0 — 25%; Group II, 26% — 39%; Group III, 40% — 59%; and Group IV, 60% or higher. The *N*'s in each subgroup respectively were 517, 686, 726 and 310 for Non-Science and 380, 405, 378 and 320 for Science.

For each subgroup the product moment correlation between *HSA* and first-term *GPA*, the corresponding regression coefficient, and the intercept of the linear regression function were obtained. In addition, the mean vector and covariance matrix for each subgroup were obtained.

Possible differences among the regression functions for the four levels of high school-college similarity were examined for statistical significance separately for the Science and Non-Science groups by applying a test on the regression of several samples developed by Gulliksen and Wilks (1950). This test establishes criteria for testing three hypotheses regarding the regression of a dependent variable upon one or more independent variables. The three hypotheses in the order that they are tested are: (1) the standard errors of estimate for each subgroup are equal, (2) the slopes of the regression

planes for each subgroup are equal, and (3) the intercepts of the regression planes for each subgroup are equal.

Results

The means and standard deviations of *HSA* and first-term *GPA*, their product moment correlation coefficient, standard error of estimate, and number of observations for each subgroup are presented in Tables 1 (Non-Science) and 2 (Science). The overall correlations between *HSA* and college *GPA* were .40 and .41 for Non-Science and Science students respectively.

TABLE 1
Means, Standard Deviations, Correlations and Standard Errors of Estimate for the Non-Science Subgroups

	Group I N = 517		Group II N = 686		Group III N = 726		Group IV N = 310	
	HSA	GPA	HSA	GPA	HSA	GPA	HSA	GPA
\bar{X}	3.00	2.29	2.85	2.26	2.80	2.42	2.60	2.44
SD	.59	.73	.51	.74	.51	.71	.54	.70
r	.351		.415		.483		.509	
SE...	.470		.456		.389		.363	

TABLE 2
Means, Standard Deviations, Correlations and Standard Errors of Estimate for the Science Subgroups

	Group I N = 380		Group II N = 405		Group III N = 378		Group IV N = 320	
	HSA	GPA	HSA	GPA	HSA	GPA	HSA	GPA
\bar{X}	3.10	2.30	2.99	2.29	2.84	2.34	2.73	2.49
SD	.54	.81	.50	.81	.55	.78	.51	.78
r	.351		.430		.470		.520	
SE...	.577		.532		.472		.444	

One method for determining the amount of information contained in a set of predictor variables for a given criterion, is to examine the spread of criterion values about the regression surface.² The greater this spread (commonly referred to as the standard error of

² Using conditional variances to measure uncertainty in cases where the criteria differ is not completely satisfactory because of the dependence of the conditional variance upon the scale parameter of the criterion variable. In the present study, however, the same criterion was employed for all subgroups.

estimate), the less information contained in the set of predictors for the specified criterion. The first test on the equality of standard errors of estimate for the four groups of Non-Science students and for the four groups of Science students yielded χ^2 values of 11.9 and 8.0⁸ respectively. Based on this test the hypothesis that high school achievement provides the same amount of predictive information about college achievement for the four levels of high school-college similarity was rejected. Tests on the equality of the slopes and intercepts of the regression functions were not performed because of significant differences among the standard errors of estimate.

Also of interest was the slight increase in the mean college *GPA* across the four levels of the percentage going to college. The overall correlations between the percentage going to college and first-term *GPA* were .09 and .07 for Non-Science and Science students respectively. Inspection of the mean *HSA*'s in Tables 1 and 2 suggests a negative relationship between high school achievement and the measure of high school quality employed in the study. The overall correlations between *HSA* and the percentage going to college were -.21 and -.27 for the Non-Science and Science curricula, respectively.

Discussion

The present results indicate that the model most commonly employed in the prediction of academic achievement, the bivariate or *k*-variate linear regression model, is inappropriate for handling the relationship among the variables studied. Linear regression, as with other linear statistical models, e.g., factor analysis, canonical variate analysis, or discriminant analysis, assumes that the variables employed have a multivariate normal distribution.

The assumption of a multivariate normal density implies linearity in the parameters and a condition perhaps too often ignored, of additivity of effects, i.e., the covariance matrix for any subset of the variables employed is a constant function of the remaining variables of the total set. Whenever the regression of the criterion upon one or more predictors is not independent of some one or more other predictor variables, as was the case in the present study, this condition of additivity is not met, and the use of the linear regression

⁸ $\chi^2_{.05} = 7.8$; $\chi^2_{.01} = 11.3$ for 3 degrees of freedom.

model is not strictly appropriate (Lee, 1961). The writers use the term "strictly" to suggest that in some cases the violation of this assumption may be so slight, i.e., the estimable interaction term is so small in a practical sense, that the linear regression model yields a tolerable approximation to the joint regression surface (cf. Lubin, 1961 for a discussion of ordinal and disordinal interaction).

Where significant interaction effects obtain, the use of the linear regression model yields, what may be called, an "averaged" surface which may be quite different from the actual or observed regression surface. The predicted criterion values generated by this model do not, in effect, "fit" anyone.

Several methods have been proposed for handling a regression situation where the criterion is a joint function of the predictors (Ezekiel and Fox, 1959; Horst, 1954; Lubin and Osburn, 1957; Saunders, 1956). Characteristic of these methods is the introduction of estimable interaction terms into the regression function; thus the criterion is expressed as a polynomial function in the predictors. In the present case, prediction of college academic achievement would be improved by expressing *GPA* as a joint function of *HSA* and percentage going to college rather than as a linear combination of these individual predictors. Despite this refinement in the prediction of the criterion, one is still left with the question of the uniformity of errors about this joint regression surface.

Again, the present results indicate that errors in prediction are not uniform over the entire regression surface. For various regions of this two-dimensional surface the error variance changes as a function of the variables employed. In terms of the hypothesis established, *HSA* does not provide the same amount of information regarding college achievement for all *Ss*; the lower is the percentage of students going to college from a given high school the less information is contained in *HSA*. Even though the accuracy of prediction is improved (through the partialling out of the estimable interaction term(s) from the error variance) through using a joint regression function, the same error distribution cannot be applied to all *Ss*. In his discussion of differential predictability and reliability, Ghiselli (1963) surveyed this problem at length and suggested that these effects are best handled by breaking down a given subject pool into more homogeneous subsets. In the present case, homogeneity is with respect to the error variance. Ghiselli's suggestion for

handling heterogeneous error variances implies separate prediction equations for each subset.

In summary, the user or investigator of *HSA* in the prediction of college academic performance is faced with two problems. The first concerns the nonadditive effects of *HSA* and high school quality in the prediction of *GPA*. In general, this problem relates to the selection of an appropriate functional form to describe the system of variables being studied. Satisfactory resolution lies in the use of statistical models which account for higher degree and interaction effects. The second problem concerns the lack of homogeneity of the error variance. This lack of homogeneity has more serious consequences for applied prediction, e.g., admissions or counseling, since the predictors do not provide uniform information about college achievement for all *Ss*. Decision rules are difficult to formulate because predictions on *Ss* from different regions of the variable space are not comparable. Breaking the total subject pool into homogeneous subsets is both costly and not completely satisfactory in that the error variance is quite probably a continuous function of the predictors.

Solution to these problems should be framed in terms of both the requirements of the user and the required decisions based on predicted values. In terms of the costs involved, the decisions to be made and the magnitude of tolerable error regarding these decisions, it may well be that the linear regression model, as was pointed out earlier, provides an adequate representation. Whenever the amount of error is intolerable with respect to these considerations, the user should adopt a strategy which reflects the subtle properties of the system of variables studied.

REFERENCES

- Bloom, B. S. and Peters, F. R. *The Use of Academic Prediction Scales for Counseling and Selecting College Entrants*. New York: Free Press of Glencoe, 1961.
- Chauncey, H. and Frederiksen, N. "The Functions of Measurement in Educational Placement." In Lindquist, E. F. (Editor) *Educational Measurement*. Washington: American Council on Education, 1951, 85-116.
- Ezekiel, M. and Fox, K. A. *Methods of Correlation and Regression Analysis*. New York: Wiley, 1959.
- Fishman, J. A. "Some Social-Psychological Theory for Selecting and Guiding College Students." In Sanford, N. (Ed.) *The American College*. New York: Wiley, 1962, 666-689.

- Ghiselli, E. E. "Moderating Effects and Differential Reliability and Validity." *Journal of Applied Psychology*, XXXVII (1963), 81-86.
- Gulliksen, H. and Wilks, S. S. "Regression Tests for Several Samples." *Psychometrika*, XV (1950), 91-114.
- Horst, P. "Pattern Analysis and Configural Scoring." *Journal of Clinical Psychology*, X (1954), 3-11.
- Laughlin, J. W. "College First Semester Academic Achievement as Related to Characteristics of a High School Graduating Class." Unpublished doctoral dissertation, The Pennsylvania State University, 1961.
- Lee, Marilyn C. "Interactions, Configurations, and Nonadditive Models." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXI (1961), 797-805.
- Lindquist, E. F. "An Evaluation of a Technique for Scaling High School Grades to Improve Prediction of College Success." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 623-646.
- Lindsay, C. A., Marks, E. and Hamel, L. S. "Academic Performance, Ability, and Attrition of Native and Transfer Students Over a Four-Year Period." *Student Affairs Research Report* 64-6. The Pennsylvania State University, 1964.
- Lubin, A. "The Interpretation of Significant Interaction." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXI (1961), 807-817.
- Lubin, A. and Osburn, H. G. "A Theory of Pattern Analysis for the Prediction of a Quantitative Criterion." *Psychometrika*, XXII (1957), 63-73.
- McClelland, W. *Selection for Secondary Education*. London: University of London Press, 1942.
- Saunders, D. R. "Moderator Variables in Prediction." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVI (1956), 209-222.
- Segel, D. *Prediction of Success in College*. Bulletin No. 15, U. S. Office of Education. Washington: U. S. Government Printing Office, 1934.
- Travers, R. M. W. "Significant Research on the Prediction of Academic Success." In Donahue, Wilma T., Coomb, C. H. and Travers, R. M. W. *The Measurement of Student Adjustment and Achievement*. Ann Arbor: University of Michigan Press, 1949, 147-190.

DEPARTMENTAL DIFFERENTIALS IN THE PREDICTIVE VALIDITY OF THE GRADUATE RECORD EXAMINATION APTITUDE TESTS¹

GEORGE F. MADAUS
Worcester State College

AND

JOHN J. WALSH
Boston College

ALTHOUGH the predictive validity of the *Graduate Record Examination Aptitude Tests (GRE-AT)* has been studied before, both the number of published studies is small and nearly all sample sizes have been rather limited. Most studies not only have grouped graduate students together, but also have ignored the variable of the major department in which the graduate students are enrolled. This study is concerned with an investigation of differences in the predictive efficiency of the *GRE-AT* for various departments in the graduate school of a New England university.

Several studies have dealt with samples composed of Ph.D. students. King and Besco (1960) found significant relationships between the *GRE-Verbal* scores and success as a research fellow. The criteria were faculty ratings and grade point averages. The *N* was 119, one of the larger samples utilized in research on the predictive validity of the *Graduate Record Examinations*. Using a sample of 46 doctoral students, Law (1960) found significant multiple *R*'s between the criterion and the *GRE-Verbal*, *GRE-Quantitative*, and the *GRE-Achievement Tests*.

¹ This work was done in part at the M.I.T. Computation Center, Cambridge, Massachusetts. Acknowledgment is made to Albert E. Beaton, Jr. of Harvard University for the development of the computer program for the IBM 7090 which was used in this study.

Michael, Jones, and Gibbons (1960) studied the relationship between scores on the *GRE-AT* and a student's standing on a background examination in chemistry at the University of Southern California. In finding a significant but slight relationship for a sample of 41 students, they concluded that the *GRE-AT* could not be used in place of their own objective background examination in chemistry. The relationship studied did not bear directly on the predictive capacity of the *GRE-AT* relative to academic success.

With a sample of 50 psychology students at the University of Florida, Robertson and Nielsen (1961) found that the correlation between *GRE-AT* scores and faculty ratings of success was statistically significant, though predictively weak from a practical standpoint.

Lannholm (1963) summarized selected studies in his ETS monograph and presented correlations between the criterion of success in Graduate School and the *GRE-AT* scores obtained from six separate studies. The magnitude of the coefficients differed from one field to another as well as within the same field between different institutions. The *N*'s varied from 16 to 96 with most *N*'s in the 20 to 40 range. Most relationships were positive and significant.

The present study utilizes data obtained from *GRE-AT* scores administered under the Institutional Testing Program during the years 1961 to 1963. All subjects were beginning graduate students at this institution. The total number in the sample was 569.

The dependent criterion variable adopted for this study was grade point average at the end of one semester of graduate study. Grade point averages were computed by assigning numerical values to graduate school grades in accordance with the following schema: A = 9, A- = 8, B+ = 7, B = 6, B- = 5, C+ = 4, C = 3, D = 2, F = 0. The number of credit hours was multiplied by the numerical equivalent of the grade received, and these weighted totals were divided by the total number of semester hours credit.

The data were subjected to an inductive multiple regression analysis. The program which was used in this study instructed the computer to select from the independent variables those which contributed to the multiple correlation in the order of their importance. Under this system, the computer first chooses that variable which correlates most highly with the criterion and then proceeds in order to select the predictor with the next highest correlation with the first

variable partialled out, and then the predictor showing the next highest correlation with the first and second variables partialled out, and so on. In cases with only two independent variables, as in this study, the program also tests to see whether the addition of the second variable to the inductively chosen first one significantly increases the predictive power of the regression equation based on the first variable alone.

A general description of the level of functioning of the sample on the *Verbal* (V) and *Quantitative* (Q) portions of the *GRE* and in terms of *GPA* is provided by the data in Table 1.

TABLE 1

Means and Standard Deviations of Predictors and Grade Point Average for 569 Beginning Graduate Students

Variable	\bar{X}	SD
GRE-V	520.08	111.26
GRE-Q	485.43	121.55
GPA	6.26	1.78

Zero order product moment correlation coefficients were computed between each of the two predictor variables and the grade point average. These coefficients are presented in Table 2.

TABLE 2

Correlations among Graduate Record Examination Scores and Grade Point Average for 569 Beginning Graduate Students

	GRE-V	GRE-Q
GPA	.19**	.18**
GRE-V		.45**

** Significant beyond the .01 level

All of the correlations in Table 2 are significantly different from a zero correlation at the .01 level.

A multiple *R* between *GPA* and the two predictor variables was computed to be .22. This is also significantly different from an *R* of zero at the .01 level.

Although all the correlations, zero order and multiple, are significant, they are, nevertheless, low. These low correlations may be due in part to the restricted spread of talent in the sample, to the fallibility of the predictors and *GPA*, and to differences between

departments. These relatively low coefficients are consistent with those reported in earlier studies.

The R^2 statistic, or coefficient of multiple determination, relates the percentage of variance in *GPA* accounted for by *GRE-V* and *GRE-Q* taken together with double consideration of common elements eliminated. In this study 4.84 percent of the variance in *GPA* was accounted for by the *GRE-V* and *GRE-Q* taken together. The remaining 95.16 percent of the *GPA* variance was explained in terms of other factors not included in this study. Some of this variance could be attributed to differences in marking and in admissions practices between departments. Other factors probably include differences in student motivation, in previous training, and in personality factors.

The multiple regression equation utilizing both *GRE* scores is the best equation possible for this grouping. The predicted grade point average is obtained by substituting the *GRE* scores of a student into the following formula:

$$\text{Predicted } GPA = .002242 (GRE-V) + .001725 (GRE-Q) + 4.255760.$$

Standard error of estimate for this equation is 1.7371 grade points.

TABLE 3

Summary Table of Pertinent Statistics, by Departments, for the Predictive Validity of the Graduate Record Examination

Department(s) (1)	N (2)	r or R (3)	Beta Weights		Constant	St. Error of Est.
			<i>GRE-V</i> (4)	<i>GRE-Q</i> (5)		
All Students	569	.22	.002242	.001725	4.255760	1.7371
Chemistry	21	.44	.009547		-.514050	2.3649
Education	181	.26	.004808		3.821579	1.8236
Economics	44	.40	.005057		4.183647	1.1882
History	28	.47	.009255		.141596	1.6479
Mathematics Institute (NSF)	68	.69	.004767	.003138	2.319321	0.7752
Nursing	42	.33		.006418	3.873051	1.3856
Mathematics, Mathe- matics Institute	86	.46	.004814		4.185731	1.0977

In Table 3 the pertinent statistics for the total sample and for the several departments within the Graduate School are summarized. For those departments for which beta weights are shown in both

columns 4 and 5, the correlation reported in column 3 is a multiple R . If only one of the columns contains a beta weight, the correlation reported in column 3 is a zero order product moment correlation coefficient. In these latter cases addition of the second GRE score did not significantly improve the prediction based on the single GRE variable.

Ten of the analyses resulted in correlations which were not significantly different from zero. These were: Biology ($N = 22$), English ($N = 68$), Mathematics ($N = 18$), Modern Languages ($N = 18$), Philosophy ($N = 12$), Physics ($N = 13$), and combinations of History, Government and Social Sciences ($N = 44$), Physics and Geophysics ($N = 19$), Classics and Philosophy ($N = 24$), and Biology and Chemistry ($N = 43$). Six of these ten analyses are of separate departments taken by themselves rather than in combination with other departments. The median N for these six departments was 18. As shown in Table 3 six other departments considered separately yielded significant correlations between the GRE scores and GPA . However, the median N for these latter six departments was 43. It would appear, therefore, that the size of N is a definite factor relative to whether or not a significant relationship is found between the dependent and the independent variables.

The findings of this study lead one to the conclusion that the practice of grouping departments for predictive purposes should not be employed. No matter how logical the grouping appears to be, the results are likely to be of limited utility. Only two exceptions are revealed by the data in Table 3. The first, the mathematics department grouped with the mathematics institute, showed significant results. However, in this case the same faculty members assigned the grades in both categories. The criterion measure GPA was, therefore, not contaminated by differences in grading practices which existed between departments. The second instance was the grouping of all graduate students as a unit. This grouping served to give a global picture of the effectiveness of the GRE scales for predicting GPA in graduate study.

The data of Table 3 indicate that GRE scores, when subjected to regression analysis, are inefficient predictors of success in this graduate school. The correlations between the GRE variables and GPA were, when the departments were taken singly and had relatively large N 's, statistically significant. The coefficients of correlation

were of magnitude similar to those reported in other studies of success in graduate work which utilized the *GRE* as the independent variable. However, from the practical standpoint, such coefficients were inefficient in forecasting power.

It appears that when *GRE* data are used with the regression model, graduate school administrators will not obtain information which will be too helpful a guide in decisions regarding admissions. This limitation is not intended to suggest that the *GRE* has no predictive validity or cannot supply administrators with useful data to guide decisions. It does, however, indicate that regression analysis is not a highly appropriate road to useful information. Alternate methods of analysis such as in the form of expectancy tables are probably more appropriate in most situations.

REFERENCES

- King, Donald C. and Besco, Robert O. "The Graduate Record Examination as a Selection Device for Graduate Research Fellows." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XX (1960), 853-858.
- Lannholm, Gerald V. "Use of Graduate Record Examinations in Appraising Graduate Study Candidates." *Graduate Record Examinations Special Report*, 1962-63. Princeton: Educational Testing Service, 1963.
- Law, Alexander. "The Prediction of Ratings of Students in a Doctoral Training Program." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XX (1960), 847-851.
- Michael, William B., Jones, Robert A., and Gibbons, Billie D. "The Prediction of Success in Graduate Work in Chemistry from Scores on the Graduate Record Examination." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XX (1960), 859-861.
- Robertson, Malcolm and Nielsen, Winnifred. "The Graduate Record Examination and Selection of Graduate Students." *American Psychologist*, XVI (1961), 648-650.

PERSONALITY RIGIDITY OF STUDENTS SHOWING CONSISTENT DISCREPANCIES BETWEEN INSTRUCTOR GRADES AND TERM-END EXAMINATION GRADES

RODNEY T. HARTNETT AND CLIFFORD T. STEWART¹

University of South Florida

At the University of South Florida, grades in all courses in the College of Basic Studies are a result of an equal combination of instructor evaluation and student performances on a common final examination. Both examination and instructor grades are originally assigned on a 15 point scale where 15 = A+ and 1 = F-, and letter grades are determined after the two distributions are combined. Although the correlation between these two measures is generally quite high, it has been recognized that a substantial number of students consistently receive a higher (or lower) evaluation on one of these two criteria.

A growing body of research (Neel, 1959; Lehmann and Dressel, 1962; Rust and Ryan, 1953; Ehrlich, 1961) suggests that personality differences may play a crucial role in determining student achievement levels in colleges. Such findings encouraged the writers to examine the phenomenon of consistent discrepancies between instructor and examination grades from the standpoint of personality, more specifically, personality rigidity.

In many respects, this study is a partial replication of research reported by Kelly (1958). Taking cues from his findings, the writers felt that students who consistently received higher grades from their instructors than from the final examination were probably those who went out of their way to make a favorable classroom im-

¹ Now at the Claremont Graduate School and University Center in Claremont, California.

pression. The writers therefore hypothesized, like Kelly, that (1) students who generally receive the higher grades from their instructors are more compulsive, conforming, and rigid than are students who generally receive the higher grades from the term-end examination, and (2) their general academic aptitude scores are lower.

Method

From a population of 1,094 students who had completed three or more Basic Studies courses, two groups were chosen: those whose final examination grades were consistently higher than their instructor grades ($N = 144$), and those whose instructor grades were consistently higher than their examination grades ($N = 153$). These two groups consisted only of those students whose mean examination grade-instructor grade differences placed them at least one and one half standard deviations from the mean difference for the entire population. This yielded a sample of 297 students with rather clear-cut consistencies favoring either the examination or instructor grade.

Students were then compared on three criteria: (1) actual mean instructor and examination grades, (2) general academic aptitude, and (3) personality rigidity. Data for their grades were available from records. General academic aptitude was measured by means of the Florida Twelfth Grade Test, a battery of tests used for admissions purposes by state universities in the state of Florida. This battery consists of the following five standardized instruments:

1. SRA Tests of Educational Ability, grades 9-12
2. Cooperative English Tests—Reading Comprehension
3. ITED—Natural Sciences
4. ITED—Understanding Basic Social Concepts
5. STEP—Mathematics

Scores on these tests are transformed to percentile ranks and the five percentile ranks are summed. One's total score, as a result, may range from 0 — 495. Because of the large number of students involved, the questionable procedure of summing percentiles has been demonstrated to be a reasonable approach.

Finally, data regarding personality rigidity were gathered by means of the *Inventory of Beliefs*, an instrument whose characteristics are described in detail in other references (Lehmann and

Dressel, 1962; Ikenberry, 1960; Dressel and Mayhew, 1954). Briefly, it is a 120 item inventory consisting of pseudo-rational statements to which the subject expresses agreement or disagreement. High scorers are considered as flexible and/or adaptive whereas low scorers might be thought of as rigid, compulsive, or authoritarian (Dressel and Mayhew, 1954). A *t*-test for uncorrelated means was used to examine the likelihood of the obtained differences occurring by chance.

Results

Table 1 indicates that students whose examination grades are consistently higher than their instructor grades have a mean examination index of 8.83 and a mean instructor index of 5.97. Students whose examination grades are consistently lower than their instructor grades have a mean examination index of 6.48 and mean instructor index of 8.88. The differences between the instructor grade means and between examination grade means are both significant beyond the .01 level.

TABLE 1
*Means, Standard Deviations and t-test Data for
Instructor and Examination Grades*

	Students whose examination grades are consistently higher than their instructor grades	Students whose instructor grades are consistently higher than their examination grades	<i>t</i>
Mean Examination Grade*	8.83	6.48	8.97**
Standard Deviation	2.29	2.44	
Mean Instructor Grade	5.97	8.88	11.11**
Standard Deviation	2.20	2.02	

* On 15-point scale where 15 = A+ and 1 = F-

** $P < .01$

In Table 2 it can be seen that students who received higher examination than instructor grades had a mean score of 361 on the Florida Twelfth Grade Test, as compared to 269 for students who received higher instructor than examination grades. Relative to the Florida Twelfth Grade Tests as a measure of aptitude, the group receiving higher examination scores would be considered to be the more able students.

Table 3 presents the *Inventory of Beliefs* findings in summary

TABLE 2

*Means, Standard Deviations and t-test Data for
Florida 12th Grade Test Scores*

	Students whose examination grades are consistently higher than their instructor grades	Students whose instructor grades are consistently higher than their examination grades	t
Mean F.T.G.T. Score	361.48	268.78	8.24*
Standard Deviation	85.10	88.82	

* $P < .01$

form. Students with higher examination than instructor grades have a mean *Inventory of Beliefs* score of 64.21, whereas students receiving higher instructor than examination grades have an *Inventory of Beliefs* mean of 60.46. As indicated in the table, the obtained difference is significant beyond the .05 level.

TABLE 3

*Means, Standard Deviations and t-test Data for
Inventory of Beliefs*

	Students whose examination grades are consistently higher than their instructor grades	Students whose instructor grades are consistently higher than their examination grades	t
Mean Inventory of Beliefs Score	64.21	60.46	2.12*
Standard Deviation	13.44	16.86	

* $P < .05$

Discussion

The findings of this study support the data resulting from a similar investigation after which this study was patterned (Kelly, 1958). Though differing in several ways, the two studies are similar enough so that the present one can be loosely termed a "partial replication" of the former.

The data reported here will lend more support to Kelly's suggestion that students who receive higher grades from their instructor than they do on final examinations are more conforming and rigid whereas students who receive higher examination than instructor grades tend to be more flexible and adaptive.

Why this is the case is not clearly apparent. Several explanations

are plausible. It can be argued (and the writers hypothesized) that the students who consistently receive higher instructor grades are those who make concerted efforts to impress their instructors and therefore influence what is largely a subjective evaluation. That these students have less aptitude than the higher examination group was demonstrated by comparison of *ACE* scores in Kelly's research, and by comparison of Florida Twelfth Grade Test scores in the present study.

On the other hand this rigidity or inflexibility may not be assisting their higher instructor rating as much as it penalizes their examination performances. That is, their consistently lower examination scores may be attributable to personality rigidity.

In any event, one is led to conclude that in situations where final grades are assigned on the basis of a combination of instructor and examination evaluation, students who consistently receive lower examination than instructor grades are more rigid and have less academic aptitude than students receiving higher examination grades.

REFERENCES

- Dressel, P. L. and Mayhew, L. B. *General Education: Explorations in Evaluation*. Washington, D. C.: American Council on Education, 1954.
- Ehrlich, H. J. "Dogmatism and Learning." *Journal of Abnormal and Social Psychology*, XLII (1961), 148-149.
- Ikenberry, S. O. "A Multivariate Analysis of the Relationships of Academic Aptitude, Social Background, Attitudes and Values to Collegiate Persistence." Unpublished Ph.D. Dissertation, Michigan State University, 1960.
- Kelly, E. G. "A Study of Consistent Discrepancies Between Instructor Grades and Term-End Examination Grades." *Journal of Educational Psychology*, XLIX (1958), 328-334.
- Lehmann, I. J. and Dressel, P. L. *Critical Thinking, Attitudes, and Values in Higher Education*. Cooperative Research Project 590, Office of Education, U. S. Department of Health, Education and Welfare. Michigan State University, 1962.
- Neel, A. F. "The Relationship of Authoritarian Personality to Learning: F-Scale Scores Compared to Classroom Performance." *Journal of Educational Psychology*, L (1959), 195-199.
- Rust, R. M. and Ryan, F. J. "Relationships of Some Rorschach Variables to Academic Behavior." *Journal of Personality*, XXI (1953), 441-456.

APTITUDE, PERSONALITY, AND ACHIEVEMENT IN SIX COLLEGE CURRICULA¹

NORMAN M. CHANSKY

North Carolina State of the
University of North Carolina at Raleigh

PROMINENT among the assumptions which have steered the development of prognosticating academic success in college are the following: That scores on ability tests satisfactorily discriminate between the academically successful student and the failure; that high school students with histories of good scholastic achievement go on to receive high marks in college; and that the mathematical average of marks, the grade point average (*GPA*) is the only adequate metric criterion of academic achievement. Vestibular research has made the first two assumptions credible. Representative of these studies was that of Anderson and Stegman (1954). They found scores on the *American Council on Education Examination (ACE)* correlated $+ .50$ with Freshman *GPA*; scores on the *Barrett-Ryan English Test* correlated $+ .56$ with *GPA*. Different sets of predictor variables were observed by Stone (1954) to explain different proportions of *GPA* in different courses of study. Specifically in the commercial curriculum, the *ACE* plus the high school grade point average (*HSGPA*) correlated with Freshman *GPA*, $+ .66$; in the elementary education curriculum the same predictors correlated $+ .73$ with *GPA*. To obtain a correlation of $+ .73$ in the physical science majors, the literature and science scores of the *Cooperative General Culture Tests (CGCT)* were added to *ACE* and *HSGPA*. The best set of predictors for social science majors was *HSGPA*, *ACE* and the

¹ This research was supported by a Grant from the North Carolina State of the University of North Carolina at Raleigh, Faculty Research and Professional Development Fund. The author is grateful for their support.

science score of the *CGCT*. The multiple correlation with *GPA*, however, was only .51.

A portion of Freshman *GPA* may, then, be attributed to ability. As much as 75 percent of the variance was not due to ability in the Anderson and Stegman (1954) study. Previous achievements also explain a portion of *GPA*. From Stone's study (1954) we learn that for social science majors as much as 75 percent of the variance in *GPA* could not be attributed to *HSGPA*, science scores on the *CGCT*, and *ACE*. The coefficients of nondetermination were lowest for elementary education and science majors, namely .47. Since decisions like student tenure in college and receipt of financial aid are based on *GPA*, greater consideration is due the limited forecasting power of the typical ability and achievement predictors.

Some workers like Munroe (1942) have sought nonintellective variables to explain achievement. She found *Rorschach* indicators of adjustment to be linked to academic success. Attempts to study behavior through using the traditional *Rorschach* scoring systems have met with obstacles. Distributions of *Rorschach* scores are often skewed, and the familiar statistical analyses cannot be easily performed. Many of the unique features of the *Rorschach* are preserved in the *Structured Objective Rorschach Test (SORT)* developed by Stone (1958). In addition, not only are all scores normalized, but their transformations are based on means and standard deviations of a nonclinical population. Freshman *GPA*, the *SORT* manual indicated, was found to be correlated with structuring scores, *F* and *F* —; anxiety, *Fch*; and responsiveness, *P*.

Admission to North Carolina State of the University of North Carolina at Raleigh (NCSUNCR) is based on a formula derived from high school rank (HSR) and the College Entrance Examination Board *Scholastic Aptitude Test Verbal Score (SAT-V)* and mathematical Score (*SAT-M*). A student is admitted when his predicted grade point average is close to 2.00 or *C*. High attrition due to dropout or transfer attests to the inadequacy of the formula in deciding upon admission.

The questions to be investigated in the present study concern the relationships of ability, high school achievement, *Rorschach* attributes and College Freshman *GPA* in each of six different curricula on the NCSUNCR campus.

Method

The deans of the several colleges on the *NCSUNCR* campus were requested to cooperate in the present study. Permission was granted to administer the *SORT* to representative samples of the 1963 Freshman class and to obtain from school records the *SAT-V*, *SAT-M*, *HSR* and *GPA*. The sample consisted of 47 students in the School of Agriculture (AG), 151 from the School of Engineering (ENG), 74 from the School of Education (ED), 96 from the School of Physical Science and Applied Mathematics (PSAM), 71 from the School of Forestry (FTRY), and 46 from the School of Textiles (TEXT).

Results

Inspection of Table 1 reveals that the original sample shrank somewhat in each school. Because no *GPA*'s were available, no attempt was made to compare stayins with dropouts. The loss, undoubtedly, does affect slightly the magnitude of the correlations.

Most *SORT* scores were observed to be uncorrelated with *GPA*. Where correlations were significantly different from zero, they were generally modest. In addition, a *SORT* variable observed to be related to *GPA* in one school, was not observed to be related to it in others. In *TEXT FM*, perception of animal movement correlated +.30 with *GPA* but in the other schools, no such relationship was observed. Also observed was an inverse relationship between *GPA* and lability, *CF* (color and poor form perception). In addition, *TEXT*, *AG* and *ED* students who were anxious, high *Fch* (perceivers of shading), received low *GPA*'s. Perception of animals (*A*), is considered a *Rorschach* attribute signifying immaturity. In *PSAM*, high *A* was related to low *GPA*. Perception of humans (*H*) was related to high *GPA* in *PSAM* but to low *GPA* in *AG*. Personality attributes were not found to be related to *GPA* in either *FTRY* or *ENG*.

It is quite evident that *Rorschach* scores, in and of themselves, do not explain academic achievement in any consistent manner. Certain personality attributes were, however, found to be slightly related to academic performance. The direction of the relationship was not always the same.

Note should be taken of the apparent differences between Stone's

TABLE 1
Zero Order Correlations of Predictors with GPA

NCSUNCR Colleges:							
<i>N</i>	Stone Study 966	AG 41	ED 68	TEXT 42	ENG 124	PSAM 83	FTRY 64
<i>Predictors</i>							
<i>SORT</i>							
Whole blot (<i>W</i>)	32	-15	+02	-11	-15	02	-15
Major details (<i>D</i>)	-28	-18	+22	+06	+15	-10	+01
Minor details (<i>Dd</i>)	09	+03	+07	-02	+04	+09	+22
White space (<i>S</i>)	04	-02	-14	00	-14	+02	-08
Form (<i>F</i>)	42	+06	+12	+09	+19	+02	-02
Poor Form (<i>F-</i>)	-46	+13	+15	+02	-10	+04	+11
Human movement (<i>M</i>)	11	-03	+01	+26	+01	+06	+18
Animal move- ment (<i>FM</i>)	-08	00	+18	+30*	+01	+04	+09
Form and color (<i>FC</i>)	20	-04	+15	+02	-01	-18	-04
Poor form and color (<i>CF</i>)	-14	+27	-32**	-28	+01	+11	+01
Shading (<i>Fch</i>)	38	-40*	-29*	-27*	-16	-09	-07
Animal Fig. (<i>A</i>)	-22	+18	+10	-03	+01	-24*	+14
Human Fig. (<i>H</i>)	33	-32*	+10	+19	+07	+33**	-03
Modal resp. (<i>P</i>)	42	+05	-13	-06	-13	-18	+01
Rare resp. (<i>O</i>)	-28	+21	+22	+13	-01	+15	+01
<i>SAT</i>							
Verbal (<i>V</i>)		+20	+24*	+38**	+16	+51**	+38**
Mathematical (<i>M</i>)		+17	+38**	+47**	+26**	+35**	+29*
Achievement							
<i>HSR</i>		+33*	+03	+26	+45**	+59**	+50**

* Significantly different from

* Significantly different from zero at five per cent level

** Significantly different from zero at one per cent level

(1954) data and the data obtained at NCSUNCR. Not only were certain *Rorschach* attributes like *P*, *E*, *F*—unrelated to *GPA* in any of the six schools on campus, but *Fch* and *O* were related to *GPA* in a direction opposite from that observed by Stone. One might speculate not only that certain universities do require more manifest ability than others but also that they may unwittingly attract or reward by means of grades students who differ with regard to personality attributes.

When the aptitude predictors are next considered, it was observed that *SAT-V* was slightly related to *GPA* in *ED*, *TEXT*, and *FTRY*. It was moderately related to it in *PSAM*. In *AG* and *ENG*, however, it was not related to it at all. The *SAT-M* was not related to *GPA* in *AG* either. The *SAT-M* was only slightly related to *GPA* in the other schools. One might have conjectured that the more

mathematically able would achieve well in engineering and the physical sciences. The obtained correlations fail to support this belief.

Based on the normalizing of achievement rank in high school when size of school is considered, *HSR* is an artifact of measurement. This statistic was not found to be related to *GPA* in ED and in TEXT. A slight relationship was observed in AG. Moderate correlations were obtained in ENG, FTRY and PSAM. Despite its being an artifact, *HSR* shows promise in predicting grade averages in the Freshman first semester, at least in some schools.

Personality, aptitude, and high school rank, then alone do not correlate with *GPA*. Yet certain scores do correlate with *GPA* in some schools. The question now raised is this: if the several variables which do correlate with *GPA* are added together, will the correlation improve? In AG combining *Rorschach* anxiety with *HSR* resulted in a multiple correlation of .49. By combining the anxiety score with the human perception *Rorschach* score, the multiple correlation with *GPA* was .45 or almost as high as the anxiety and high

TABLE 2
Multiple Correlations of Predictors with GPA

School	Predictors	R
Agriculture	Fch, H	.45
	Fch, HSR	.49
	Fch, H, HSR	.49
Education	CF, Math	.48
	Fch, Math	.43
	CF, Fch, Math	.51
Textiles	FM, Verbal	.42
	FM, Math	.51
	Fch, Verbal	.43
	Fch, Math	.51
	Verbal, Math	.45
	FM, Verbal, Math	.53
Engineering	Math, HSR	.49
PSAM	H, Verbal	.56
	H, Math	.45
	H, HSR	.63
	Verbal, Math	.54
	Verbal, HSR	.66
	Math, HSR	.65
Forestry	H, Verbal, HSR	.69
	Verbal, HSR	.58
	Math, HSR	.52
	Verbal, Math, HSR	.58

school rank combination. At most, the sets explain less than 25 percent of the variance in *GPA*. For this school, high anxiety plus low high school rank and high anxiety plus aversion for perception of humans are related to academic attainment. Adding *Fch* to *H* and *HSR* did not improve the correlation at all. (See Table 2.)

In ED, mathematical aptitude combined with the absence of emotional lability explained slightly less than 25 percent of the variance in *GPA*. The multiple *R* obtained was .48. Freedom from anxiety and aptitude for mathematics explained 17 percent of the variance in *GPA*. The multiple *R* was .43. The combination of variables did explain *GPA* better than did any single variable; yet the correlations were not too high. Adding *Fch* to *CF* and *SAT-M*, moreover, did not improve the correlation to any great extent.

Multiple correlations in TEXT were moderate, too. Animal movement plus verbal aptitude did explain *GPA* as well as did the *Rorschach* anxiety score combined with verbal aptitude and the mathematics aptitude combined with verbal aptitude. These sets explained 17 to 20 percent of the *GPA* variance. Mathematical aptitude contributed much to the *GPA* and in combination with animal movement and anxiety explained about 25 percent of the variance. When the animal movement scores were added to both *SAT-V* and *SAT-M*, 28 percent of the variance was explained.

In ENG, the mathematics aptitude scores and *HSR* correlated .49 with *GPA*. This set does not improve the prediction of academic success over the *HSR* alone. High achievers in ENG, then, have tended to obtain good marks in high school.

In PSAM, the sets of predictors observed were the most promising. Since *HSR* plus perception of human figures correlated +.63 with *GPA*, 38 percent of the variance was explained. The human figure perception plus the verbal aptitude correlated + .56 with *GPA*. This result was as high as the correlation of verbal and mathematics aptitude scores with *GPA*. The correlations of *HSR* with *GPA* was .59, slightly higher than that of other predictors. Adding *SAT-V* and *H* to *HSR* improves the correlation slightly. Alone, *HSR* explains 35 percent of *GPA*; together with the other two it explains 48 percent.

The *HSR* together with *SAT-M* correlated .52 with *GPA* in FTRY. Together with *SAT-V*, *HSR* correlated .58 with *GPA*. The correlation, then, improves only slightly. Adding *SAT-V* and *SAT-*

M to *HSR* does not produce any higher correlation with *GPA* than does *SAT-V* and *HSR* together.

Conclusions

The fact that the several courses of study are oriented on somewhat different cognitive domains, it is hypothesized, may in part explain the lack of agreement in results. No all-or-none promise of an overriding importance of one or two predictor variables could be found. In no school did the commonly used aptitude and achievement variables always correlate to a high degree with *GPA*. Neither did personality variables consistently explain *GPA* in all schools. It appears quite clear to the writer that whatever it takes to obtain a particular *GPA* in any one course of study is insufficient to explain it in another. Although the inherent weaknesses of the *GPA* may be attenuating the relationship between the predictor and criterion variables, the predictor variables studied did offer some clue about the direction in which future research might go. The predictors currently being used are relevant, but too confining to permit judicious admissions decisions. More must be learned about the student traits which are linked to grades received as well as about the factors considered by instructors who give the grades. The hypothesis concerning differences in personality characteristics of the students attracted to the several courses of study requires examination. Equally important is the hypothesis that certain combinations of personality, ability and achievement will be approved of in certain college environments but not in all. Uncovering those sets which fit in which schools and in which curricula will improve the admissions decisions greatly. The viability of the *GPA* is at stake as a criterion. Examined elsewhere (Chansky, 1964) is its inherent grouping error as well as the lack of validity and reliability of grades. Affecting the *GPA* is the fact that students drawn from such populations as agriculture, engineering, education, and textiles which differ in initial abilities, interests, and achievements are welded together to shape distributions of grades. Adding to its speciousness as a statistic, *GPA* is insensitive to the differences in achievement between the student whose *GPA* is 3.00 but is based on a heavy load of physics, zoology and economics courses and the one whose *GPA* is 3.00 but is based on a light load of first aid, visual aids, and farm machinery courses.

A deeper philosophical issue, however, is at stake. Higher education has set goals to develop self realization, economic efficiency, and responsible citizenship in its charges. Yet these less tangible but perhaps ultimately more important objectives await investigations.

Summary

Correlated with *GPA* were measures of anxiety, human figure perception, and high school rank in the School of Agriculture; poor form perception, anxiety, and mathematics aptitude in the School of Education; animal perception, poor form perception, anxiety, verbal, and mathematics aptitude in the School of Textiles; mathematics aptitude and high school rank in the School of Engineering; animal figure perception, human figure perception, verbal aptitude, mathematics aptitude, and high school rank in the School of Physical Sciences; and verbal aptitude, mathematics aptitude, but especially high school rank in the School of Forestry.

REFERENCES

- Anderson, M. R. and Stegman, E. J. "Predictors of Freshman Achievement at Fort Hays Kansas State College." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XIV (1954), 722-723.
- Chansky, N. M. "A Note on the Grade Point Average in Research." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXIV (1964), 95-99.
- Munroe, R. L. "An Experiment in Large-Scale Testing by a Modification of the Rorschach Method." *Journal of Psychology*, XIII (1942), 229-263.
- Stone, J. B. "Differential Prediction of Academic Success at Brigham Young University." *Journal of Applied Psychology*, XXXVIII (1954), 109-110.
- Stone, J. B. *The S-O-Rorschach Test*. Monterey, California: California Test Bureau, 1958.

VALIDITIES AND INTERCORRELATIONS OF MMPI SUB- SCALES PREDICTIVE OF COLLEGE ACHIEVEMENT¹

PHILIP HIMELSTEIN

Texas Western College, El Paso

THE value of a personality questionnaire which is uncorrelated with measures of intellectual ability but is correlated with school achievement is obvious. Presumably, such an instrument would account for a portion of the variation in school performance unaccounted for by measures of scholastic aptitude. In a multiple regression equation, the noncognitive scale would materially increase the predictive efficiency of college entrance examinations. There have been many efforts to devise and validate such an instrument. For example, Kleinmuntz, (1962) lists eleven scales developed from the pool of items in the *MMPI* which were developed to predict academic success among college populations. One such scale, containing 26 *MMPI* items, was constructed by Altus (1948) and was found to be correlated .39 with grade-point average and .21 (nonsignificant) with a measure of intelligence, the *Altus Measure of Verbal Aptitude*.

The purpose of the present paper is to determine the intercorrelations among *MMPI* subscales selected from among those described by Kleinmuntz. At the same time, the study will attempt to determine the relationship of each subscale to a measure of scholastic achievement (grade-point average) and to scholastic aptitude (American College Test, or *ACT*). This study can be viewed as a combination of predictive validity and postdictive validity, since the administration of the noncognitive predictors occurs between the

¹ These data were gathered during the author's tenure at New Mexico State University and the study was facilitated by a grant from the New Mexico State University Research Center.

administration of the measure of scholastic aptitude and the accumulation of the grade-point average.

Procedure

The subjects were drawn from sections of introductory psychology at New Mexico State University. These subjects were tested in small groups with a booklet which combined the items of seven of the eleven subscales listed by Kleinmuntz. These subtests are:

- Ac*, Academic Achievement (18 items)
- Ae*, College Achievement (26 items)
- Gr*, Graduate School Potential (14 items)
- Hr*, Honor Point Ratio (16 items)
- Ie*, Intellectual Efficiency (39 items)
- Or*, Originality (25 items)
- Un*, Underachievement (22 items)²

This phase, which provides the raw data for the intercorrelational study, included 281 students. Since only entering freshmen are tested with the American College Test (*ACT*), moderate shrinkage in sample size occurred for that aspect of the study concerned with correlations between the various subtests and the measure of scholastic aptitude ($N = 193$). One full semester after the administration of the *MMPI* items, the grade-point average of each subject was calculated.

Results

The intercorrelations among the *MMPI* subscales and the measures of academic achievement and aptitude are summarized in Table 1. Also included in this table (below the diagonal) is the number of *MMPI* items in common between pairs of the subtests. Thus an index of overlap was provided.

From Table 1, we can observe that all of the subscales, with the exception of *Un*, have significant correlations with grade-point average. Five of the seven scales have correlations with academic achievement significant beyond the .01 level of probability, whereas one subtest is significant at the .05 level. It would appear, then, that these subscales have validity as predictors of success in school.

² The reader is referred to the article by Kleinmuntz (1962) for the references to the original sources for these subtests.

TABLE 1

Intercorrelations among MMPI Subscales with Each Other and with Academic Aptitude and Achievement

Subtest	Ac	Ae	Gr	Hr	Ie	Or	Un	ACT	GPA
Ac		180	469	545	589	133	-114	404	370
Ae	1		281	300	243	051	-165	217	158
Gr	2	1		413	651	131	-118	305	284
Hr	2	0	1		575	304	054	441	387
Ie	6	2	2	6		107	-071	373	342
Or	0	(1)	2	1	0		041	241	240
Un	(1)	(1)	1	0	1	(1)		-143	-034
ACT									596

Decimal points have been omitted.

Figures below the diagonal are the number of overlapping items in two subtests. Numbers in parentheses are the number of items scored in the opposite direction.

For all columns except ACT, $p = .05$ for r of .118; $p = .01$ for r of .177. For the ACT column, r 's required are .138 and .181, respectively.

Examination of Column Eight of Table 1, which summarizes the correlations between the seven subscales and the ACT, provides an unexpected finding. All of the subscales are significantly correlated with the measure of scholastic aptitude. The *Un* scale, an instrument designed to predict college failure rather than college success, is negatively correlated with the ACT. For all subscales the correlation with the measure of scholastic aptitude is slightly greater than is the correlation of these subtests to a measure of scholastic achievement. This is not the first reported finding of a relationship between an MMPI subscale and an intellectual variable. Gough (1953) obtained an r of .22 between his *Hr* scale and the *Altus Measure of Verbal Aptitude*, which for his sample of 104 college students, would be significant at the .05 level. It is interesting to note that in his table of correlations, for seven non-university samples (high school and military officers), the relationship is nonsignificant.

The results of the present study strongly imply that the noncognitive predictors of school performance employed may not be independent of intellectual factors and may be, in reality, indirect measures of intelligence.

REFERENCES

- Altus, W. D. "A College Achiever and Nonachiever Scale for the MMPI." *Journal of Applied Psychology*, XXXII (1948), 385-397.
- Gough, H. G. "The Construction of a Personality Scale to Predict

Academic Achievement." *Journal of Applied Psychology*, XXXVII (1953), 361-366.

Kleinmuntz, B. "Annotated Bibliography of MMPI Research among College Populations." *Journal of Counseling Psychology*, IX (1962), 373-396.

CONCURRENT VALIDITY OF THE TEST OF ENGLISH AS A FOREIGN LANGUAGE FOR A GROUP OF FOREIGN STU- DENTS AT AN AMERICAN UNIVERSITY

HENRY DIZNEY
Kent State University

THE purpose of this study was to investigate the concurrent validity of the *Test of English as a Foreign Language (TOEFL)* for a group of foreign students at Kent State University. *TOEFL* is an Educational Testing Service test which became available for use in 1964. The establishment of *TOEFL* test centers in the home countries of foreign students is an important administrative advantage over other English proficiency tests. That fact, rather than technical merit, could encourage premature institutional acceptance of the test.

Validity evidence, as is the case with most English proficiency tests for foreign students, is scanty. An ETS memorandum concerning *TOEFL* indicates that a sample of 512 foreign students seeking admission to U. S. colleges who took the test on February 17, 1964 had a mean total score of 500 and standard deviation of 85. Estimates of reliability of scores are unavailable, but correlations between total *TOEFL* scores and institutional rankings on English proficiency at Columbia University, New York University, and the University of Michigan provide some validity data. For these three universities, given in non-identifiable order, the correlations between *TOEFL* total scores and English proficiency ratings were .78, .87, and .76 with *n*'s of 215, 91, and 45, respectively.

Procedures

In September, 1964, 20 foreign students at Kent State took *TOEFL*, the *Michigan Test of English Language Proficiency*, and

the *American College Testing Program* (ACT) English subtest. In addition, each student had been rated on English proficiency by the foreign student advisor.

The Measures. TOEFL is a three and one-half hour test divided into five parts. Although scores are available for each subtest, only total TOEFL converted scores were used in this study. The *Michigan Test* is a 75-minute examination having three parts. Total *Michigan Test* raw scores were used in this study. The ACT English test is a 50-minute subtest from the larger battery. It is designed as a usage test for American students entering college. For purposes of this study, ACT English raw scores were used.

The Michigan Test was developed by the English Language Institute of the University of Michigan in 1962. Its manual estimates the reliability by a corrected split-half method to be .965. The validity of a parent test was estimated to be .51 by cosine approximation to tetrachoric r for academic performance dichotomized as satisfactory-unsatisfactory.

Earlier editions of ACT have indicated the reliability for the English subtest to be .84 by correction of a split-half estimate. The correlation between ACT English standard scores and English grades for one year was .565 for a recent Kent State sample of 1699 freshmen.

The ratings for English proficiency by the foreign student advisor were obtained six weeks after the fall quarter began. It was felt that after this period of time, which included an orientation program and considerable student contact, the advisor would have sufficient familiarity with the foreign students to enable him to meaningfully rate them. The advisor was simply asked to rank order the foreign students according to his opinion of their overall proficiency in English. This was done with no test information available to the advisor.

The Sample. The twenty foreign students, 19 male and 1 female, were all new students to Kent State. They came from 15 countries. Thirteen had been in the United States for less than one month and 17 for less than one year; the maximum length of stay was three years. They reported having had from five to 17 years of English instruction with the median being nine years. Sixteen academic major fields were reported. Their ages ranged from 19 to 38 with a median of 23.5 years.

Results

The means and standard deviations, respectively, for each of the three tests were: *TOEFL* (converted standard scores for total test) 538.10 and 106.19; *Michigan Test* (raw score units for total test) 79.35 and 17.94; *ACT English* (raw score units) 30.70 and 14.44. The Pearson r 's among each of the objective measures taken by the sample are reported in Table 1.

TABLE 1
Correlations between *TOEFL*, *Michigan Test*,
and *ACT English Scores*, $N = 20$

	TOEFL	ACT-English
Michigan Test	.97*	.65*
TOEFL		.74*

* Significant at .01 level

Rank order correlation coefficients between the foreign student advisor's proficiency rankings and ranks on each of the three objective tests yielded the following results: *Michigan Test* .78, *TOEFL* .75, and *ACT-English* .62.

Conclusion

Within the situational limitations of this study, an exceptional degree of congruence is evident between *TOEFL* and the *Michigan Test* ($r = .97$). The use of both of these tests for assessing the English proficiency of foreign students may, therefore, represent wasteful redundancy. In terms of concurrent validity, *TOEFL* "accounts" for almost 13 percent more *ACT-English* variance than does the *Michigan Test*. The difference between *TOEFL* and the *Michigan Test* in terms of agreement with the foreign student advisor's ratings, although favoring the latter test, is negligible. The use of other criteria, e.g., grade-point averages, was precluded in this study by the fact that the subjects lacked comparability in terms of curricula and instructional levels.

Additional studies involving both *TOEFL* and the *Michigan Test* are planned as more data become available with time. These studies should undertake the assessment of predictive validity for varying curricula and differing characteristics of the examinees.

SELECTION TECHNIQUES FOR PAKISTANI POSTGRADUATE STUDENTS OF BUSINESS¹

GLEN GRIMSLEY

University of Southern California

GEORGE W. SUMMERS

Business Administration, University of Illinois

THIS study has three major purposes. The first is to make an initial report on the effectiveness of techniques and tests used in selecting Pakistani graduate students for a Master of Business Administration degree program. The second purpose is to present comparative norms of United States and Pakistani students on the selection tests used. Finally, the study serves as an illustration of the application of discriminant analysis. This type of analysis was developed to find optimal weights for independent variables for situations wherein the criterion is categorical rather than quantitative. Discriminant analysis makes it possible to perform significance tests and to make population estimates on the basis of normal curve assumptions. It thus represents both an extension of the power of most multiple cutting score techniques and an alternative thereto.

The locale and focus of the work reported here is the Institute of Business Administration, University of Karachi, Karachi, Pakistan. The Institute is modeled after colleges of business in the United States. It offers a two-year study program leading to an *MBA* degree. The curriculum for the first year is fixed. All classes are conducted in English.

Students of the Institute come from all parts of Pakistan. In February faculty teams go to designated cities to test and to interview applicants for entry the following August. Applicants from Karachi are processed at the Institute in June.

¹ The work reported here was done while the authors were serving as advisors to the Institute of Business Administration, University of Karachi, Karachi, Pakistan, under a University of Southern California Graduate School of Business Administration technical assistance contract with the United States Agency for International Development.

To be considered, an applicant must have received a bachelor's degree from a recognized institution. An applicant who has attended full time throughout his previous schooling can have completed all work necessary for a bachelor's degree in fourteen years. Typically, the study of English was begun in the sixth grade, and English was used as the medium of instruction beginning in the tenth grade.

Selection of students for the Institute is comprised of three major steps. All applicants must take one of two timed test batteries. After elimination of applicants who score below a predetermined cut-off point, faculty teams conduct structured individual interviews with the remainder. The final step is taken at meetings of the Selection Committee in late June or early July. About 150 entry permits are issued to those selected from approximately 400 eligible applicants. Of these 150, from 100 to 120 students actually appear for classes in August.

The IBAT Battery

The Institute of Business Test Battery (*IBAT*), which was constructed by one of the authors, consists of five tests. These are English vocabulary (64 items—10 minutes), English sentence structure (20 items—7 minutes), arithmetic (40 items—3 minutes), numerical reasoning (25 items—6 minutes), and symbolic reasoning (30 items—5 minutes). The first, fourth, and fifth tests are multiple choice in format.

The SCAT Battery

Forms 1A and 1B of the *Cooperative School and College Ability Test (SCAT)* were designed by the Educational Testing Service for use with high school seniors, college freshmen, and sophomores in the United States. Each form is composed of 4 parts. Part IV contains several items which require the testee to do computations which involve United States coins.² These items were rewritten to reflect the names of Pakistani monetary denominations rather than those of the United States.

² Part IV of the quantitative section of Forms 1A and 1B of the *SCAT* administered to *IBA* students was modified to reflect the names of Pakistani currency denominations, rather than those of United States currency denominations.

Comparisons: Pakistani and United States Testees

The *IBAT* was first made useful for the selection committee by establishing percentiles for each of the five tests. The scores used were those of the first 98 students who attended the Institute for at least one month and who took the tests during the normal selection process prior to their admission and attendance in either August, 1961, or August, 1962. Comparative scores were obtained from a group of United States students. The group was composed of 102 students in the Graduate School of Business Administration at the University of Southern California. The tests were given after their admission early in their first semester of work on *MBA* requirements.

Comparative results are also available on the *SCAT* battery. The Pakistani percentiles reported on this battery were obtained by giving the battery to students attending the Institute on a specific day in February, 1964. Results are reported in Table 1 for 136 full-time day degree students. Sixty-seven took Form 1A of the test, and sixty-nine took Form 1B. The raw score conversions used were the same as those given in the *SCAT* administration manual. The United States percentiles were obtained by slight simplification of results available from the publishers of the *SCAT* battery. Those which appear in Table 1 are from the fall testing of 1134 students in grade 14 of 97 U. S. colleges. The results from the *IBAT* were substantially in agreement with those shown in Table 1.

Optimal Selection with the IBAT Battery

Two techniques offered themselves as alternatives for weighting the five tests in the *IBAT* battery. Multiple correlation of test scores and average grades was one possibility. Discriminant analysis was the other possibility. The most important consideration in choosing between them is that a student cannot begin a new semester if he has failed more than one subject in the previous semester. Thus, the major criterion of performance is categorical—a situation for which discriminant analysis was designed.

One form of discriminant analysis assumes that the two statistical populations to one of which an item will be assigned are multivariate normal with identical dispersion matrices. This form of discriminant analysis also assumes that any significant regression

TABLE 1

Percentiles for U. S. and Pakistani Students on the SCAT Battery

Percentile	Converted Test Scores			
	Verbal		Quantitative	
	IBA	United States	IBA ¹	United States
99	327	335	334	338
90	308	320	326	327
80	302	313	321	321
75	300	310	317	317
70	299	307	316	315
60	295	303	311	311
50	293	300	308	307
40	290	295	304	302
30	288	292	300	296
25	287	291	298	294
20	284	286	296	288
10	277	276	291	277
1	264	256	279	256

relationships among variables within either population are linear.³

Two types of check were performed with respect to the normality assumption. Frequency polygons were plotted on the five performance test scores within the success group and, separately, within the failure group. In addition, cumulative plots were made on normal probability paper. Apparent conformity with a normal curve assumption was found for vocabulary, sentence structure, and numerical reasoning scores. A square-root transformation of symbolic reasoning scores was necessary to remove positive skewness. The success-group distribution of arithmetic test scores showed that too much time had been allowed for the 40 items. Since this distribution was truncated at a point only slightly above the median score, no normalizing transformation was possible. Rather than discard this variable immediately, however, another approach was taken.

Vocabulary scores within the success group had a mean of 39.86. The mean vocabulary score for the failure group was 35.55, and t was 2.13 (162 degrees of freedom) for the difference between these means. The geometric means of students' vocabulary and arithmetic scores were found to be essentially normally distributed within the success group and within the failure group. The means of this new variable were 35.39 for the success group and 32.79 for the

³ Kendall, M. G., *A Course in Multivariate Analysis*, New York, Hafner Publishing Company, 1957, 144-170.

failure group. The t value for the difference in these means was 2.46 with 162 degrees of freedom. This result led to substitution of the geometric mean of vocabulary and arithmetic test scores for vocabulary score alone.

The check on the assumption that any regression within either population is linear was performed by making scattergrams of all pairs of test scores within each population. No evidences of curvilinearity were found.

Within the set of successes and separately within the set of failures, individuals were cross-classified by the class with which they entered and by their place of residence. Tests for homogeneity of variance and differences among mean scores on the five performance tests of the battery disclosed that the special class which began in October, 1962, was atypical. Discarding it from the analysis left 164 cases which met the assumptions.

Now that the data which appear to comply with the assumptions of the selected form of discriminant analysis have been isolated, the inputs to the analysis can be described formally. An eligible case for the study is a student who went through the standard selection procedure. The student also had to attend the Institute long enough to have an appraisal of his performance. This was taken to be a grade in all courses for the first compulsory monthly examinations. Eligible cases were divided into success and failure groups. Successes completed their first semester without failing more than one course. Thus they were scholastically qualified to enter the second semester. Failures were not qualified to enter the second semester by reason of dropout attributable to poor scholastic performance or to failure in more than one course for their first semester.

The characteristics of the data remaining in the study can be shown in a table. Four variables are defined: X_1 (geometric mean of vocabulary and arithmetic scores), X_2 (sentence structure score), X_3 (numerical reasoning score), X_4 (square root of symbolic reasoning score). Two additional definitions are \bar{X}_i , the sample mean, and s_i^2 , the unbiased sample estimate of variance, for the i^{th} variable. Thus, the summary is as shown in Table 2.

F tests on the 4 variance ratios between successes and failures showed no significant differences. The t tests on the differences between success and failure means yielded $t_1 = 2.46$, $t_2 = 1.28$, $t_3 = 1.46$, and $t_4 = 0.31$. The first value is significant at the 2 percent level.

TABLE 2

Summary of Data to be Used in IBAT Discriminant

Data	Success Group	Failure Group
n	124	40
\bar{X}_1	35.39	32.79
\bar{X}_2	9.00	8.20
\bar{X}_3	11.35	10.32
\bar{X}_4	2.69	2.63
s_1^2	35.375	28.125
s_2^2	12.130	10.985
s_3^2	15.158	15.046
s_4^2	0.864	0.679

The Discriminant

The discriminant model used here has the form

$$D = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4,$$

wherein the X_i are test scores for each individual as defined in the previous section. The purpose of the analysis is to find those values of the a_i which produce maximum separation between the success and failure groups. The measure of separation is the difference in mean D scores per unit of variance in D scores. This variance is assumed to be constant within both success and failure populations.

The first step in the analysis is to calculate the sums in Table 3. The symbol x_i used in that table is defined to be a deviation from the mean, \bar{X}_i .

Division of each entry except n in the final column of Table 3 by

TABLE 3

Number of Cases, Sums of Squared Deviations, and Cross Products Measured from the Means

	Success Group	Failure Group	Combined
n	124	40	164
$\sum x_1^2$	4351.15	1096.88	5448.05
$\sum x_2^2$	1492.00	428.40	1920.40
$\sum x_3^2$	1864.39	586.78	2451.16
$\sum x_4^2$	106.28	26.49	132.78
$\sum x_1x_2$	1088.94	256.26	1345.20
$\sum x_1x_3$	936.05	209.30	1145.35
$\sum x_1x_4$	109.56	3.91	113.48
$\sum x_2x_3$	534.00	190.40	724.40
$\sum x_2x_4$	138.70	13.72	152.42
$\sum x_3x_4$	94.37	46.51	140.87

162 furnishes the pooled variance—covariance matrix. Symmetrical elements omitted, this matrix is shown as the top of Table 4.

Let I_{ij} be the element from the i^{th} row and j^{th} column of the inverse matrix. Let m_i be the difference between the mean scores on the i^{th} variable for the success and failure groups. Then the coefficients which maximize the separation between groups are

$$a_i = I_{i1}m_1 + I_{i2}m_2 + \cdots + I_{i4}m_4.$$

Substitution in the equation immediately above yields

$$\begin{aligned} a_1 &= (.0375)(2.60) - (.0225)(.80) - (.0112)(1.03) + (.0057)(.05) \\ &= .0683, \end{aligned}$$

$$a_2 = .0097, \quad a_3 = .0359, \quad a_4 = -.0444.$$

TABLE 4
Variance—Covariance Matrix and its Inverse
(Symmetrical elements omitted)

Variance—Covariance Matrix				
	V_1	V_2	V_3	V_4
V_1	33.63	8.30	7.07	0.70
V_2		11.85	4.47	0.94
V_3			15.13	0.87
V_4				0.82
Inverse of Variance—Covariance Matrix				
	I_1	I_2	I_3	I_4
I_1	0.0375	-0.0225	-0.0112	0.0057
I_2		0.1143	-0.0179	-0.0929
I_3			0.0798	-0.0545
I_4				1.3798

Thus the discriminant function is

$$D = .0683X_1 + .0097X_2 + .0359X_3 - .0444X_4.$$

Statistical Significance

A modification of the usual analysis-of-variance test for the significance of multiple regression applies here. The test involves three steps. First, the estimated variance of the D -scores within each population is obtained. This is simply the difference of the mean D -scores for the two samples.

$$\bar{D}_1 = \sum a_i \bar{X}_{1i}, \quad \bar{D}_2 = \sum a_i \bar{X}_{2i}$$

and

$$S_D^2 = \bar{D}_1 - \bar{D}_2 = 2.7907 - 2.5709 = 0.2198.$$

The second step has as its purpose the separation of the total sum of squares in the D -scores into the portion attributable to discrimination and into a portion representing residual variation. The total sum of squares is standardized to 1.00 for convenience. The calculations are

$$A = \frac{n_1 n_2}{(n_1 + n_2)^2} S_D^2 = (.1844)(.2198) = .0405$$

$$K = \frac{A}{1 + A} = \frac{.0405}{1.0405} = 0.038952$$

$$1 - K = 0.961048$$

wherein n_1 and n_2 are the sample sizes for the success and failure groups, respectively.

Finally, K and $1 - K$ serve as the standardized sums of squares from discrimination and residual variation in the analysis of variance. The degrees of freedom for discrimination is the number of input variables (p) and the degrees of freedom for the residual is $n_1 + n_2 - p - 1$. The F ratio for the mean squares is 1.61 with 4 and 159 degrees of freedom. This value is not high enough to suggest that the discriminant function will produce significant separation between success and failure population means.

A t test is also available to see which coefficients in the discriminant function are significantly different from zero. This test involves standardizing coefficients and dividing each standardized coefficient (a'_i) by the sample estimate of its standard error $S_{a'_i}$. The formulas for these two quantities are

$$a'_i = \frac{n_1 n_2}{(n_1 + n_2)^2} (1 - K) \cdot a_i$$

and

$$S_{a'_i}^2 = I_{ii} \cdot \frac{n_1 n_2}{(n_1 + n_2)^2} \cdot \frac{1 - K}{n_1 + n_2 - p - 1}$$

The results for this study are

$$t'_{a_1} = \frac{a'_1}{S_{a'_1}} = \frac{0.0121}{0.0064} = 1.89,$$

$t'_{.1} = 0.15$, $t'_{.5} = 0.68$, and $t'_{.9} = 0.20$ each with 162 degrees of freedom. None of these is significant.

Further Analyses

The t values just cited show that variables beyond X_1 contribute no additional discrimination in the four-variable case. They also indicate that X_2 is possibly the weakest member of the set. It was eliminated and a three-variable discriminant function was calculated with the same type of result. The same type of result also occurred when the best two-variable discriminator was built. Thus no combination of variables can be expected to do as well as X_1 alone in separating members of the success population from those of the failure population. Recall that $t = 2.46$ for the difference in means on this variable.

Expected Performance

The most effective combination of variables to discriminate between success and failure groups appears to be the geometric mean of vocabulary and arithmetic scores. It is expected that these X_1 scores for the success group will form a normal population with a mean of about 35.39 and a standard deviation of about 5.80, whereas the failure group will be normally distributed about 32.79 with the same standard deviation.

Based on the above estimates, the expected performance of this discriminant is shown in Table 5. This table shows that, for instance, some 32 percent of the failure group can be expected to lie below a cutting score of 30, whereas only 18 percent of the success group will lie below this cutting score.

Another aspect of expected performance is shown in Table 6. By

TABLE 5

*Estimated Proportions of Success and Failure Groups
Lying below Selected Cutting Scores*

Cutting Score	z_f	z_s	Portion of Failure Group	Portion of Success Group	Difference in Proportions
25	-1.365	-1.790	.086	.037	.049
30	-.481	-.930	.316	.176	.140
32.79	0	-.448	.500	.327	.173
34.09	.224	-.224	.589	.411	.178

comparing the relative heights of normal curve ordinates at selected values of X_1 , one may calculate an estimate of the probabilities of success and failure, given a designated X_1 score.

TABLE 6
*Estimated Probabilities of Success and Failure
for Persons Having Designated Scores*

Score	Ordinates at z				Probabilities		
	z_1	z_2	f	s	Sum	f	s
25	-1.365	-1.79	.157	.080	.237	.66	.34
30	-.481	-.930	.355	.259	.614	.58	.42
34.09	.224	-.224				.50	.50
40	1.242	.795	.185	.295	.480	.38	.62
45	2.110	1.655	.043	.101	.144	.30	.70

Summary and Discussion of Results

Two batteries of tests were used in the selection of Pakistani postgraduate students of business administration. One battery, the *IBAT*, was created especially for this purpose. Portions of this battery were found to discriminate between the group of students who succeeded and the group which failed in a master's degree program. The second battery, a modified *SCAT* was administered to the students entering classes in August, 1964. Its effectiveness as a discriminator will be reported in the future.

Although the *IBAT* discriminated between successes and failures in a statistically significant fashion, its power to discriminate was not so high as is desirable operationally. Fifty-nine percent of the success population and forty-one percent of the failure population are estimated to lie above the optimal cutting score.

Comparative norms for United States and Pakistani students indicate that the two groups are essentially equal on tests of number ability such as those used in this study. Pakistani students are, however, at a noticeable disadvantage in tests of the verbal factor in English.

THE PREDICTIVE VALIDITY OF ELEVEN TESTS AT ONE STATE COLLEGE

RICHARD W. BOYCE AND R. C. PAXSON

Troy State College (Alabama)

Background

THE predictive validity data reported in a test manual are often intended to be true or representative of a national sample. Since the discrepancies in predictive validity coefficients are likely to change as a local college sample departs from a national sample, validity studies should be performed for each college. It might be expected that at Troy State College, where the local median has ranged from the twenty-sixth percentile nationally to the thirty-eighth percentile on any of several tests over a five year period, local validation studies would be appropriate and meaningful for colleges like Troy.

Problem

The purpose of this study was to determine estimates of the local predictive validity of various standardized aptitude tests which have been used in admissions and guidance programs as well as of tests on several non-cognitive variables.

Description of Predictor Variables

The tests used were:

1. *The American College Testing Program Examination (ACT)*
2. *California Test of Mental Maturity, 1963 Revision (CTMM)*
3. *California Capacity Questionnaire (CCQ)*
4. *American Council on Education Psychological Examination for College Freshmen (ACE)*

5. *Cooperative School and College Ability Tests (SCAT)*
6. *College Entrance Examination Board Scholastic Aptitude Test (SAT)*
7. *College Qualification Test (CQT)*
8. Rokeach's Dogmatism Scale (*D*)
9. California F-Scale (*F*)
10. Cree Questionnaire (*Cree*)
11. *California Achievement Tests, 1957 Edition (CAT)*

The tests were administered from 1959 to 1964. The *ACE* and *SCAT* were administered during freshman orientation. The *ACT* and *SAT* were given during the senior year in high school. The *CAT* was administered during the student's eleventh year in school. The other tests were administered during the first quarter in college. Because of the high drop-out rate, the first quarter in college has the least restriction of range.

Samples

The tests were administered to random samples of 100 freshman students who had been admitted to the college and who were in the college general education program. All students are in this general education curriculum, except for a few who transfer to other colleges at a later time. More than half of the students in the general education program become teachers after graduation.

Criterion Measures

The criterion measure used was the grade point average at the end of the student's first quarter in college. This measure was employed because the drop-out rate is very high after the first quarter. (A more or less open door policy was used until 1964-65.) Every student in the general education program is required to take the same courses during the first two years. Reliability estimates of the grading practices have usually ranged in the .80's and .90's. A college regulation demands a normal distribution of grades, and grade distribution charts are made each quarter by professors and are published in a report to the entire college faculty. The grades are converted to a point average where $A = 4$, $B = 3$, $C = 2$, $D = 1$ (lowest passing grade) and $F = 0$ quality points. The average student load is 15 hours per quarter.

Results

Table 1 shows correlations between predictor and criterion variable of the grade point average (GPA) ($N = 100$).

TABLE 1
Correlations between Predictor and Criterion Variable

Predictor	Correlation	Year Administered
1. <i>ACT</i>		
English	.64**	1960, 1961,
Mathematics	.47**	1962, 1963,
Social Studies	.50**	1964, 1965
Natural Science	.46**	
Composite	.57**	
2. <i>CTMM</i>		
Language	.62**	1964
Non-Language	.56**	
Total	.64**	
3. <i>CCQ</i> (total)	.52**	1964
4. <i>ACE</i> (total)	.42**	1959
5. <i>SCAT</i>		
Verbal	.49**	1959, 1960,
Quantitative	.42**	1961, 1962,
Total	.46**	1963
6. <i>SAT</i>		
Verbal	.36**	1962, 1963
Quantitative	.38**	
Total	.46**	
7. <i>CQT</i> (total)	.44**	1962
8. <i>D</i>	.29**	1964
9. <i>F</i>	.24*	1964
10. <i>Cres</i>	-.24*	1964
11. <i>CAT</i>	.57**	1960

* Significant at the .05 level.

** Significant at the .01 level.

Table 2 shows the intercorrelation between the high school grades and the college overall GPA.

TABLE 2
Intercorrelation of High School Grades and College Overall GPA
 $N = 777$

High School Grades	(2)	(3)	(4)	(5)	Regression Weight	\bar{X}	σ
(1) English	0.486	0.585	0.511	0.444	0.193	2.49	0.97
(2) Mathematics		0.429	0.479	0.350	0.085	2.22	1.07
(3) Social Studies			0.495	0.362	0.076	2.68	0.93
(4) Natural Science				0.324	0.041	2.34	0.96
(5) College Overall GPA						1.20	0.66

Multiple Correlation = 0.482
Standard error of estimate = 0.578
Regression Constant = 0.235

Table 3 shows the distribution and percentile ranks of high school grades in this group.

TABLE 3
Distribution and Percentile Ranks (PR) of High School Grades
N = 777

Grade	English		Mathematics		Social Studies		Natural Science	
	F	PR	F	PR	F	PR	F	PR
A	130	92	100	94	165	89	100	94
B	247	67	206	74	279	61	227	73
C	278	34	275	43	263	26	301	39
D	114	8	157	15	63	5	139	10
F	8	1	39	3	7	1	10	1

Table 4 shows the intercorrelations of *ACT* scores and the College *GPA*.

TABLE 4
Intercorrelation of ACT Scores and College Overall GPA
N = 777

ACT	(2)	(3)	(4)	(5)	Regression Weight	\bar{X}	σ
(1) English	0.429	0.543	0.437	0.314	0.028	17.4	4.41
(2) Mathematics		0.451	0.550	0.238	0.011	17.3	5.34
(3) Social Studies			0.625	0.314	0.025	17.9	5.14
(4) Natural Science				0.214	-0.005	18.0	5.26
(5) College Overall GPA						1.2	0.66

Multiple Correlation = 0.365
Standard error of estimate = 0.614
Regression Constant = 0.165

Cross-Validation Data

Random samples of 50 freshmen other than in the validity study were drawn over the years on all tests except for the *CTMM* which used another 1964 sample. Tests on the differences of the means and *F*-tests indicated that the students over the years have been very much alike on the variables measured. The correlations in the cross-validation samples were similar to those of the validity samples. Studies at other Alabama institutions have shown similar results.

The *CTMM*, probably, cannot be compared to previous editions, because the 1963 revision replaced the ratio IQ with the deviation IQ (which, of course, is not an IQ, but a standard score).

The positive relation between the *F*-scale and academic aptitude

is not consistent with that found in most validity studies at other colleges across the nation. The "closed environmental matrix" from which most of the student body is drawn probably accounts for the high *F* scores. The high positive relation between the *D* and *F* furnished evidence of the restricted belief system. The inverse relation between the *F* and *Cree* has been consistently found from freshman to graduate levels in other local studies. One possible conclusion is that prejudice in a group resulting from enculturation operates in a manner different from the pathological type and does not necessarily depress cognitive functioning. The negative correlations between measures of creativity and dogmatism indicated that both the tests and grades are tapping the more convergent rather than divergent abilities.

THE PREDICTIVE VALIDITY OF THE SCHOOL COLLEGE ABILITY TEST (SCAT) AND THE AMERICAN COLLEGE TEST (ACT) AT A LIBERAL ARTS COLLEGE FOR WOMEN

PAUL A. DE SENA
John Carroll University

AND

LOUISE ANN WEBER
Hoban-Dominican High School

Introduction

SINCE the inauguration of the *School and College Ability Test* (SCAT) in 1955, numerous studies have been conducted to test its ability to predict scholastic achievement in college. As of the present time, very little research has been reported in which the *American College Test* (ACT) of 1959 has been correlated with collegiate scholastic achievement to determine its predictive validity. Because the ACT is currently utilized as a criterion for admission by many universities, it is imperative that its predictive qualities be statistically validated.

Problem

It was the purpose of this study (1) to find the degree of correlation between the Verbal, Quantitative, and Total scores of the *School and College Ability Test* (SCAT) and grade-point averages of the students from one class that had completed two semesters of work at Notre Dame College; (2) to ascertain the extent of correlation between the English, Mathematics, Social Studies, Natural Sciences, and Composite scores of the *American College Test* (ACT) and grade-point averages of students from another class that had completed two semesters of work at Notre Dame College; and (3)

to determine which of these two aptitude tests had the higher degree of predictability of successful college achievement.

Procedures

The converted scores for each part of the *SCAT* were obtained for 77 students of the 1958 class of Notre Dame College. Percentiles for each part of the *ACT* were determined for 92 students of the 1960 class of Notre Dame College. The populations for both classes included all of the students, excluding the religious, who had completed two semesters of work at the college. Grade-point averages for two semesters of college work were calculated for both classes.

Since one part of the study data was reported by means of converted scores (*SCAT*), and the other in percentiles (*ACT*), it was appropriate to use rank correlation and then to convert to estimated coefficients of product-moment correlation.

Coefficients of rank correlation were converted to Pearson's coefficients of correlation. This step was necessary in order to estimate whether there was a significant difference between the coefficients. The level of confidence was tested for each coefficient of correlation. Coefficients of correlation calculated between *SCAT* Total scores and the grade-point averages and between *ACT* Composite percentiles and the grade-point averages were converted to Fisher *z*-scores.

The significance of the difference between the two coefficients of correlation was computed by means of the following formula:

$$\sigma_{D_r} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

where N_1 is the size of one sample and where N_2 is the size of the second sample.

The critical ratio was calculated by dividing the difference between the *z*-coefficients by the standard error of the difference between the two *z*-coefficients. After the critical ratio was obtained, a conclusion was formulated.

Findings

Coefficients of rank correlation calculated between each part of the *SCAT* and grade-point averages and between each part of *ACT*

and grade-point averages of the two classes of Notre Dame College students are given in Table 1. Rank coefficients converted to coefficients of correlation are also included in the table.

TABLE 1
*Correlations between Test Scores and Grade-Point
Averages After Two Semesters of College*

Test	Coefficient of Rank Correlation	Rank Coefficients Converted To Correlation Coefficients
<i>ACT</i> English	0.39	0.41*
<i>ACT</i> Mathematics	0.42	0.44*
<i>ACT</i> Natural Sciences	0.39	0.41*
<i>ACT</i> Social Studies	0.34	0.35*
<i>ACT</i> Composite	0.50	0.52*
<i>SCAT</i> Verbal	0.58	0.60*
<i>SCAT</i> Quantitative	0.60	0.62*
<i>SCAT</i> Total	0.65	0.67*

* Significant beyond the 0.01 level of confidence.

Since the numerical value for the correlation coefficient for the *SCAT* and grade-point averages was greater than the correlation coefficient for the *ACT* and grade-point averages, it was necessary to determine whether the former correlation was significantly larger than the latter. After the method of determining the reliability of the difference between the correlations (0.67 for the *SCAT* Total and 0.52 for the *ACT* Composite) was applied, it was found that the difference between the two coefficients of correlation was not significant.

Conclusion

Based on the evidence presented, it can be stated that the coefficient of correlation between the *SCAT* Total and grade-point averages is not necessarily higher than is the coefficient of correlation for the *ACT* Composite scores and grade-point averages when two different samples of subjects were studied. Although there is not a significant difference between the two correlations, the observed difference suggests that the *SCAT* Total may possibly be more predictive of college achievement than is the *ACT* Composite Score.

VALIDATION OF THE KAHN INTELLIGENCE TESTS¹

ERNEST D. McDANIEL

AND

WILLIAM T. CARSE

University of Kentucky

CROSS-CULTURAL developmental studies and comparative education projects continue to suggest the need for measures of mental development which are relatively free of cultural and educational contaminants. Within the United States accurate measures of intelligence of the culturally deprived, foreign born, and minority groups have been hampered by the high loading of middle-class verbal material characteristic of most tests. Davis and Eells (1953), Cattell (no date), Lieter (1940), Raven (1938), Goodenough (1926), and others have made well known contributions in this area. However, new instruments which hold promise of being less culture-bound and more flexible in application than existing ones warrant further exploration and development. The *Kahn Intelligence Tests: Experimental Form (KIT:EXP)* appears to be such an instrument (Kahn 1960).

The *Kahn Intelligence Tests* is composed of a felt strip and of sixteen plastic objects including butterflies, dogs, hearts, stars, a cross, a parrot, an anchor, a circle, and a segment of a circle. The nature of the tasks presented to the subject may be illustrated by an item from the five-year old level. Three plastic dogs are placed in a row so that the small black dog is facing left, the small white dog is standing in the middle, and the large black dog is facing right. The two black dogs are lying flat on the table. After the arrange-

¹ The research reported herein was supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education and Welfare.

ment has been exposed to the child for five seconds, the objects are covered with the felt strip for 5 more seconds; then these objects are handed to the child with instructions to recreate the original order. A score of two months of mental age is given if there are no errors in position or in facing. There are six items for each year of mental age from one through fourteen.

Kahn reports that the suggested age norms are based on 40 adults and 297 children. Further descriptions of the normative group are not provided. A test-retest reliability coefficient of .94 is reported for 23 children ranging in age from 1 to 14. An estimate of validity was obtained for the same 23 children by correlating mental age scores from the *Kahn* with scores from the 1937 *Revised Stanford-Binet Intelligence Scale* ($r = .75$). Both of these correlation coefficients must be viewed with caution. Even a crude index of development is likely to correlate highly with mental age measures over such a wide age range. Thus, in the evaluation of the *Kahn Intelligence Tests*, the most obvious need is for a replication of the validity study within a more restricted age range and with a larger sample of children.

Procedure

Fifty children (25 boys and 25 girls) between the ages of four years two months, and six years ten months, were tested with both the *Kahn* and the *Revised Stanford-Binet Intelligence Scale, Form L-M*. The children were all urban and predominantly of managerial or professional parents.

Results

The correlation coefficients between the *Kahn* and the *Binet* were .83 when computed on the basis of IQ scores, and .85 when computed with mental age scores. Both coefficients are significant, i.e., above the .28 required for 48 degrees of freedom. More importantly, the coefficients are sufficiently high to offer encouragement that the *Kahn* does provide scores similar to those obtained on the *Binet*.

The mean intelligence quotient and mental age obtained on each test are presented below:

Scale	Intelligence		Mental Age	
	Mean	Sigma	Mean	Sigma
<i>Stanford-Binet</i>	107.9	13.4	71.2	12.6
<i>Kahn</i>	106.4	15.6	73.2	15.6

The application of *t* tests indicates that there are no significant differences between means obtained on the two tests whether one looks at intelligence or at mental age.

Freedom from Socio-Economic Bias

Kahn claims that his test is relatively culture-free. Because of the homogeneity of the sample used in this study, only suggestive evidence can be obtained to evaluate this claim. Each of the 50 children was placed into a group designated either "middle to upper" or "middle to lower." Assignment to a group was done by the examiner on the basis of parent's occupation, educational level, place of residence, and some admittedly subjective cues noted at time of testing. The hypothesis guiding this analysis of the data is simply that if the *Binet* favors the middle class child and penalizes the lower class child and if the *Kahn* is not so biased, then for the upper socio-economic group, the mean *Stanford-Binet* IQ should exceed the mean *Kahn* IQ and that for the lower group, the mean *Kahn* IQ should exceed the mean *Stanford-Binet* IQ. The mean IQ for these two groups is presented below for each test:

	<i>Stanford-Binet</i>	<i>Kahn</i>
Upper Group (<i>n</i> = 28)	111.26	112.22
Lower Group (<i>n</i> = 22)	103.99	100.22

It may be noted that the mean intelligence quotients for a particular socio-economic group are almost identical whether they are determined by the *Binet* or the *Kahn*. Although the groups studied are not so divergent as would be desired, the difference in average intelligence between "upper" and "lower" suggests two distinct groups of children; moreover, this distinction is even slightly more apparent on the *Kahn* test than on the *Stanford-Binet* instrument.

Summary and Conclusions

The *Kahn Intelligence Tests* provide scores which are surprisingly similar to those of the *Binet* within the age range of the sample studied. Claims of freedom from cultural bias are not supported by this study, although the absence of clearly divergent socio-economic groups indicates that this latter finding must be considered only as suggestive.

Relatively easy to administer, the *Kahn Intelligence Tests* does seem to avoid a heavy loading of verbal or school-related materials. Experimental work with the *Kahn* should be continued, as the evidence presented suggests that this relatively undiscovered instrument is highly promising.

REFERENCES

- Cattell, R. B. *Handbook for the Individual or Group Culture Free Intelligence Test*, Champaign, Illinois: Institute of Personality and Ability Testing. (no date)
- Davis, A. and Eells, K. *Davis-Eells Games*, Yonkers-on-the-Hudson, New York: World Book Company, 1953.
- Goodenough, F. L. *Measurement of Intelligence by Drawings*, New York: Harcourt, Brace and World, Inc., 1926.
- Kahn, Theodore C. "*Kahn Intelligence Tests: Experimental Form (Kit)*" *Perceptual and Motor Skills*, Monograph Supplement V10-1, Missoula: Montana State University Press, 1960.
- Lieter, R. G. *The Lieter International Performance Scale*. Volume 1, Santa Barbara, California: Santa Barbara State College Press, 1940.
- Raven, J. C. *Guide to Using Colored Progressive Matrices*. London: H. H. Lewis and Company, 1938.

USES OF COGNITIVE AND NON-COGNITIVE TEST MEASURES IN SIXTY-FOUR PRIVATE LIBERAL ARTS COLLEGES: IMPLICATIONS FOR PREDICTIVE VALIDITY AND ASSESSMENT OF CHANGE

ERNEST L. BOYER

State University of New York

AND

WILLIAM B. MICHAEL

University of California, Santa Barbara

DURING the early months of 1965 a survey of college practices in the measurement of student characteristics was undertaken by the Commission on Experimentation and Research of the Council for the Advancement of Small Colleges (CASC). By early July questionnaires were returned by 64 of the 72 Council colleges, a group of institutions seeking to promote close student-teacher encounter in a context of a liberal arts program.

The questionnaire focused on the following topics: (1) citation of the aptitude, achievement, vocational interest, and personality tests administered to students prior to or at the time of admission; (2) indication of whether any one of these tests would be administered again in order to assess change in student characteristics during college years; (3) reactions of the college administration to the assessment of the characteristics of students, and (4) the propriety or impropriety of collegiate stress on the personality development of the student. Other attitudinal items called for administrator opinions concerning (1) the reaction of the board of trustees and constituency to an intensive program of college testing including efforts to assess personality characteristics, (2) the administrative location of a student evaluation program, and (3) the ways in

which data regarding student characteristics and student change should be used on campus once they have been secured.

Purpose

Thus it was the purpose of the writers to report briefly the information obtained from a questionnaire regarding the measurement of student characteristics and their changes and to suggest implications for predictive validity and assessment of change in this group of 64 colleges. Whenever appropriate, a breakdown in the findings will be reported relative to three classificatory variables: (1) size of institution (large versus small in terms of whether the enrollment was equal to or greater than 351 or less than 351), (2) church versus non-church in terms of the identification or lack of identification of the controlling board with a specific religious body, and (3) status of accreditation (accredited versus non-accredited).

The numbers of colleges belonging to each of the eight permutations of the three classificatory variables were as follows:

<i>Classification</i>	<i>Number of Colleges</i>
Church—Large—Accredited	13
Church—Large—Non-accredited	9
Church—Small—Accredited	7
Church—Small—Non-accredited	10
Non-Church—Large—Accredited	7
Non-Church—Large—Non-accredited	6
Non-Church—Small—Accredited	1
Non-Church—Small—Non-accredited	11
	<hr/> 64

Findings

The results may be organized in terms of responses concerned with (1) cognitive tests, (2) measures of vocational interest or aptitude, (3) measures of personality, and (4) additional reactions.

Cognitive Tests

The overwhelming majority of the 64 colleges in this sample made use of a scholastic aptitude test. Only 3 of the large colleges and 6 of the small colleges did not employ such an instrument as a part of their admissions program; of the 28 accredited colleges and 36 non-accredited colleges only 4 and 5 colleges, respectively, did not report that an effort was made to measure scholastic aptitude through use of a test. Although only 2 of the 39 church colleges did

not cite the aptitude test requirement, 4 of the 25 non-church colleges made no mention of such a test requirement. Virtually all the well known instruments of scholastic aptitude were cited. The *Scholastic Aptitude Test* (SAT) of the College Entrance Examination Board was the most frequently used instrument with 20 citations followed by the *American College Test* (ACT) and *School College Ability Test* (SCAT) with 13 and 8 enumerations, respectively. Although the SAT was the leading test in most of the categories of college classification, the ACT was the most popular instrument for the 19 non-accredited church colleges, with 8 reporting its use.

Although scholastic aptitude measures were frequently used, tests of knowledge or achievement were much less common. Among the 64 colleges studied, 30 reported no utilization of achievement tests. Specifically, corresponding to the 20 accredited church colleges, 19 non-accredited church colleges, 8 accredited non-church colleges, and 17 non-accredited non-church colleges the numbers of colleges not utilizing achievement tests were 9, 7, 4, and 10, respectively. A large variety of tests was named by those colleges that sought to measure achievement in discrete fields of study. Tests involving reading and language skills were the most prominent. The *Sequential Test of Educational Progress* (STEP) received the largest number of citations—nine in all.

Although slightly more than 50 percent of the colleges administered a test of knowledge to incoming freshmen, 48 out of the 64 colleges, or 75 percent, indicated that no instruments (or equivalent forms of them) were administered a second time during the four college years to assess the changes in student knowledge. Specifically, of the 20 accredited church colleges, 19 non-accredited church colleges, 8 accredited non-church colleges, and 17 non-accredited non-church colleges, the numbers not assessing for change were 15, 15, 6, and 12. In the few colleges that did retest the students considerable variety in the instruments employed was noted, although tests of communication skills were among those most commonly readministered. Gains in knowledge of the Bible were assessed in five colleges. Size of school was not related to the incidence of assessment of change. Even in those colleges in which measurement of change was sought, the evaluation programs were extremely limited in scope.

Measures of Vocational Interest

Of the 64 colleges studied, 49—more than 75 percent—made no use of vocational interest scales during the students' four college years. Thus among the 20 accredited church colleges, 19 non-accredited church colleges, 8 accredited non-church colleges, and 17 non-accredited non-church colleges, the numbers *not administering* any measures of vocational interest were 15, 13, 5, and 16. Among those colleges using such measures, the vocational interest scales prepared by Strong and by Kuder were the most commonly employed devices.

Of the colleges making use of vocational instruments only one indicated that it might possibly readminister a test of vocational interest to assess change.

Measures of Personality Characteristics

Of the 64 colleges sampled, only 19 reported use of personality measures in their student testing program. Specifically, among the 20 accredited church colleges, 19 non-accredited colleges, 8 accredited non-church colleges, and 17 non-accredited non-church colleges, the numbers *not administering* any measures of personality were 13, 15, 4, and 13. Among the colleges that sought to measure personality characteristics of students, most of the well known instruments designed to assess adjustment, temperament, attitudes, and values were cited as being used. No marked relationships between incidences of use of such instruments and college size, church or non-church classification, or accreditation status were noted.

Additional Reactions of the Respondents

In response to open end questions regarding the measuring of student characteristics at the time of admission to colleges, 54 out of the 64 college administrators responding declared that their college should know more than it did. There were no striking differences associated with accredited or non-accredited status, church or non-church classification, or size of institution. Most of the individuals suggested that they wished to see greater use of measures of personality, motivation, attitudes, and other noncognitive attributes.

Furthermore 58 of the 64 respondents expressed a desire to know

more about changes that occur within students while they are at college. Again, the highly diversified comments indicated considerable desire to assess changes not only in cognitive processes associated with curricular objectives but also in values, attitudes, and personality characteristics of students. Despite this concern, answers to previous items revealed that virtually nothing was being done to determine changes in student characteristics of either a cognitive or non-cognitive nature.

Regarding whether there might be a danger to the assessment of changes in the sense that colleges might find out too much about their students, only 18 of the replies were decidedly affirmative and 10 were equivocal or conditional. In other words, a substantial majority of institutions wanted as much information as possible. The major concerns of the minority group of respondents were that test data might be misused or misinterpreted and that other highly important objectives of the college experience might be subordinated or deemphasized in light of the results of test data, the reliability and validity of which would be open to serious question. Concerning this matter of the dangers of student assessment, there were marked differences in the replies of institutions relative to their accreditation status. Twenty-one of the twenty-six non-accredited institutions yielded affirmative or conditionally affirmative responses concerning the possible dangers related to the assessment of changes among students. On the other hand, only seven of the twenty-eight accredited institutions expressed such reservations. There was also a slight tendency for the smaller colleges in comparison with the larger colleges to express caution regarding attempts to assess changes in students.

With respect to the question concerning whether the college's board of trustees and constituency would fully support a move to gather more information about students at the time of admission and about changes that occur during the college year, 51 out of the 64 respondents suggested that these groups would support such a move.

Relative to who should be in charge of a student evaluation program on campus, virtually every conceivable college official was cited except for the president. No discernible pattern of responses could be associated with any of the major variables of classification. Although the Dean of Students received the greatest number of

nominations—16 in all—directors of counseling and guidance, directors of institutional research, academic deans, departments of education and psychology, student personnel officers, testing bureaus, cooperative teams (committees), registrars, admission officers, and others were mentioned.

As to whether a separate office for student evaluation should be established, the opinions expressed were overwhelmingly against such a proposal. Specifically, 34 opposed the suggestion, 16 gave affirmative answers, 4 submitted conditional replies, and 10 failed to answer. No noticeable pattern of responses was associated with any of the three major variables of classification.

Suggestions concerning ways in which data regarding student characteristics and student changes should be used were numerous and highly diversified. Common responses included the following purposes: (a) individual counseling on learning and adjustment problems, (b) placement of students and scheduling of their courses, (c) diagnosis of learning difficulties, (d) assessment of the objectives of the curriculum, (e) evaluation of instructional policies and procedures, (f) establishment of new instructional goals and modifications of others on the basis of research findings, and (g) mechanics and procedures for recording and disseminating test data.

Implications of Findings

Analysis of the responses to the questionnaires revealed that among private liberal arts colleges little systematic effort was being directed toward (1) the determination of the predictive validity of cognitive or non-cognitive measures administered to students prior to or at the time of admission or (2) the assessment of changes in either cognitive or non-cognitive characteristics of students. However, a careful evaluation of the reactions of the respondents to additional items regarding the merits of such assessments clearly indicated that most colleges were highly interested in and desirous of finding ways to measure changes in both cognitive and non-cognitive characteristics of students. That the responsibility for student evaluation was frequently centered in the hands of personnel who for the most part were not highly trained in educational and psychological measurement might account for the lack of follow-up testing and assessment of changes in the characteristics of college students. It would appear that even though a majority of the respondents

opposed the establishment of a special office for student evaluation, perhaps the most satisfactory approach to the evaluation of admissions procedures, of changes in student characteristics, and of the attainment of the educational objectives by private liberal arts colleges would be to centralize the effort in an office of institutional evaluation or research, which would operate within a policy determined cooperatively by representative committees of the faculty and the college administration.



BOOK REVIEWS

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

- Ebel's Measuring Educational Achievement.* LEWIS R. AIKEN, JR. 1167
- Cattell's Personality and Social Psychology.* ANDREW L. COMREY 1170
- Dubois' An Introduction to Psychological Statistics.* PETER A. TAYLOR 1174
- Guenther's Concepts of Statistical Inference.* PETER A. TAYLOR 1176
- Burket's Reduced Rank Models for Multiple Prediction.* JAMES A. WALSH 1181
- Noll's Introduction to Educational Measurement.* SAMUEL T. MAYO 1184
- McLaughlin's Interpretation of Test Results.* NORMAN C. MABERLY 1186
- Spiegel's Theory and Problems of Statistics.* JAMES M. RICHARDS, JR. 1188
- Lopez's Personnel Interviewing: Theory and Practice.* PETER F. MERENDA 1189
- Farber and Wilson's Conflict and Creativity.* JAMES M. RICHARDS, JR. 1191
- Stewart and Warnath's The Counselor and Society: A Cultural Approach.* HENRY KACZKOWSKI 1193
- Miller's Guidance Services: An Introduction.* HENRY KACZKOWSKI 1196
- Mosher, Carle, and Kehas' Guidance: An Examination.* HENRY KACZKOWSKI 1199
- Crow and Crow's Organization and Conduct of Guidance Services.* HENRY KACZKOWSKI 1200
- Lazarus' Personality and Adjustment and Rotter's Clinical Psychology.* PHILIP HIMELSTEIN 1201
- Huber's Influencing Through Argument.* JAMES M. RICHARDS, JR. 1203



Measuring Educational Achievement by Robert L. Ebel. Englewood Cliffs, N. J.: Prentice-Hall, 1965. pp. xii + 481.

For those who are awaiting an updated version of Gulliksen's *Theory of Mental Tests*, this is not it. Although this is a thorough, carefully-written book about testing, its orientation is more verbal than quantitative. The book gives a *practical* treatment of construction, administration, and scoring of teacher-made achievement tests, little attention being paid to standardized tests of aptitude, achievement, or personality. By limiting his discussion to achievement tests, the author has avoided the problem of predicting future performance (with aptitude tests) which may be more uncertain than assessing present knowledge or skill.

Dr. Ebel, currently Professor of Education at Michigan State University, is a well-known authority on testing, having gained experience as Director of the University Examinations Services at Iowa State and as Vice President for Testing Programs and Services at Educational Testing Service. The present book, which the author states is "... by no means the simplest textbook on educational measurement that is available; nor is it the most thorough" is written in a scholarly, if sometimes rather technical, style. Although it was the author's intention that the book be useful for students with no previous training in educational measurement, this will not be an easy book for beginning students to read. The instructor who adopts the book should be prepared to do much explaining and repeating, especially of the material in chapters 8, 10, and 11.

The general format of the book consists of 13 chapters, each divided into numbered sections, the headings of which pose a problem, a question, or suggest a procedure. Then the numbered summary statements at the end of a chapter draw conclusions about these problems or procedures. There is also a glossary of 150 measurement terms in an appendix, followed by an index.

Brief summary statements of the content of each of the 13 chapters follow.

Chapter 1, "The Need for Better Classroom Tests," is a general chapter concerned with the reasons teachers should study testing and the uses of classroom tests.

Chapter 2, "What Should Achievement Tests Measure?," is a

lengthy discussion of the educational objectives or goals which may be measured by achievement tests.

Chapter 3, "How to Plan a Classroom Test," classifies test questions into seven "mental process" categories and gives numerous examples of multiple-choice items in each category.

Chapter 4, "The Characteristics and Uses of Essay Tests," contrasts essay with objective tests in terms of similarities, differences, and the situations in which each is more appropriate. In addition, suggestions for preparing and grading essay tests are made.

Chapter 5, "How to Use True-False Tests," is somewhat unusual in that it is a whole chapter devoted to true-false items. It is refreshing to have some of the virtues of the true-false item noted, rather than their shortcomings simply enumerated, as is sometimes done. A good survey of the literature, numerous examples, and a discussion of the advantages and disadvantages of this type of item are presented.

Chapter 6, "How to Write Multiple-Choice Test Items," is a long "how to do it" chapter, consisting largely of illustrations of desirable and undesirable items. True-false and multiple-choice are the only two types of objective test items discussed extensively in this book, presumably because they have greater flexibility. But the reviewer holds that matching, ranking, completion, and other types of items are very appropriate in certain situations, and some discussion of the characteristics of these would have added to the usefulness of the book.

Chapter 7, "How to Administer and Score an Achievement Test," discusses factors such as test-taking skills, test anxiety, cheating, honor systems, kinds of answer sheets, scoring methods, corrections for guessing, and differential weighting systems. The reviewer found this chapter especially interesting and comprehensive. It was observed, however, that no mention was made of optical scoring methods, such as that used in the IBM 1260. Admittedly this is an unimportant omission, but it may date the book somewhat.

Chapter 8, "Describing Test Scores Statistically," shows how to arrange scores in a frequency distribution and to compute such statistics as the arithmetic mean, the median, the variance, the standard deviation, percentile ranks, stanine scores, and the correlation coefficient. Lathrop's short-cut estimate of the standard deviation, viz. dividing one-half the number of scores into the difference between the sums of the upper and lower one-sixths of the scores, should be a blessing to students who have trouble extracting a square root!

Chapter 9, "How to Judge the Quality of a Classroom Test," discusses the test qualities of relevance, balance, efficiency, objectivity, specificity, difficulty, discrimination, reliability, fairness, and unspeediness.

Chapter 10, "How to Estimate, Interpret, and Improve Test Reliability," is a rather technical coverage of the concept of reliability. Unique for an introductory text in educational measurement is the derivation of the Kuder-Richardson and Spearman-Brown formulas from a consideration of the ratio between test covariance and variance. The beginning student should anticipate some difficulty with this chapter. Also, confusion may be caused by the units on the abscissa of Figure 10.1, which are in geometric rather than arithmetic progression.

Chapter 11, "How to Improve Test Quality Through Item Analysis," is another chapter which will require the assistance of an instructor. The use of discrimination and difficulty indices in comparing item responses of the upper and lower 27 percent groups on total test score is detailed, and the effects of revising items in the light of item analysis data are discussed. It should be noted that this method of item analysis tends to create a test of homogeneous content, and although it is convenient when measuring achievement, the use of an external criterion may be a more appropriate procedure when devising a test to *predict* behavior.

Chapter 12, "The Validity of Classroom Tests," is a practical, relatively non-technical discussion of achievement test validity. The author points out that validity is a complex concept with various interpretations and that the concept is more meaningful when it refers to tasks which the test samples rather than to traits which it purports to measure. The relationship between reliability and validity is also discussed. This chapter would obviously have been more extensive and more technical if the book had dealt with tests of aptitude and personality.

Chapter 13, "Marks and Marking Systems," is a thorough discussion of the thorny problem of grade assignment. An improved, specific marking system, which takes into account the ability level of the class, is also presented.

In actuality, the reviewer finds much that is commendable and very little to criticize in this book. It should be useful as both a textbook and a reference source. Of particular significance are the reviews of the literature and the recommendations on topics such as weighting, corrections for guessing, item analysis, reliability, and validity. The book is obviously carefully written and carefully proofread, and, as stated by the author, it is practical. It tells how to construct, administer, and score tests, and it does this well. And although it is true that the book does not cover all kinds of tests and that it will require study to understand, these are qualifications which are more likely to be counted as advantages than shortcomings.

LEWIS R. AIKEN, JR.

University of North Carolina at Greensboro

Personality and Social Psychology by Raymond B. Cattell. San Diego: Robert R. Knapp, 1964. pp. x + 799.

This book consists of 62 articles selected from the more than 235 papers published during the nearly four-decade writing career of Raymond B. Cattell. This remarkable output of research and theoretical papers is in addition to his 22 books and monographs, many tests and manuals, and 22 chapters in books edited by others. To review this book is, in a sense, to review the career of the author. It would be difficult to do this without consideration of other publications by the same author which are not actually included in this volume for lack of space.

The book, which unfortunately has many proof-reading errors, is organized in such a way as to emphasize the tremendous breadth of Cattell's interests and contributions to psychology. Twelve sections are included which cover articles on various aspects of personality theory, clinical psychology, physiological psychology, social psychology, genetics, and methodology. Unlike the published writings of some psychologists with long bibliographies, Cattell's works have an extensive background of empirical research. Article after article reveals evidence of diligent data-gathering activities, painstaking scholarship, and the productive application of a highly creative mind. Some of his important contributions include: development of a major personality theory, development of a systematic set of factors for the description of human personality, formulation of the group syntality concept and a new way of approaching the study of social psychology, extensive research in the area of measuring personality by means of objective tests, important work in the definition and measurement of intelligence, unique contributions to the study of genetic versus environmental determination of ability and personality characteristics, and numerous methodological contributions, particularly in the application of factor analysis. Most of the remainder of this review will be devoted to a consideration of some possible targets for criticism in this otherwise meritorious work.

Cattell has sometimes left himself open to criticism by making inferences, claims, and extrapolations which to cautious research people seem to be unwarranted. For example, he has been criticized for claiming identity of factors in behavior rating, questionnaire, and objective test realms. He often utilizes small loadings in the identification of factors, a practice which is especially questionable when samples are rather small, as they often are in Cattell's studies. He has overwhelmed people working in the personality area with a large vocabulary of strange names and a voluminous outpouring of results which virtually defy collation. The reviewer would have been pleased to see a greater concentration of Cattell's great research efforts in a more circumscribed area, with stronger emphasis upon

replication of results, larger samples, and careful elimination of alternate possible interpretations of the data.

His published tests have not always seemed to be in line with the claims made for them. Levonian (1961), for example, carried out a statistical analysis of the 16 Personality Factor Questionnaire and found that only about 10 percent of the significant interitem correlations were between items on the same factor, that almost 25 percent of all intrafactor correlations were in a direction opposite to that intended, that about a third of the items had no significant intrafactor correlation, and that the average item correlated significantly with fewer than one other item in its factor but with nearly eight items outside its factor. The factor scales, therefore, are apparently very heterogeneous. Cattell has since taken the position that homogeneity of item content in a factor scale is not only unnecessary, but also actually undesirable. In one of the articles in this volume, he criticizes Comrey for suggesting that high homogeneity is a necessary but not sufficient condition for factor purity. He goes on to develop a formula, incorrectly printed in the text, for the correlation between a scale and a factor as a function of item homogeneity, number of items used, and the item validities. Applying this formula under various hypothetical conditions, Cattell comes to the conclusion: "It is clear that, as we argued introductively on general principles, for any given mean goodness of items, in terms of correlation with the factor, higher validity of the scale is obtained more readily, with fewer items, when homogeneity is lower" (p. 760). What Cattell does not point out, however, is that this principle is valid only if it is admitted that the factor can be complex. If so, this demonstrates nothing more than the well-known principle of multiple regression that uncorrelated but valid tests are best for predicting a complex criterion. Cattell himself, however, speaks constantly of univocal factors in his writing as the desired goal of factor analysis. Thus, we must presume that he does not really wish to admit complex factors to justify low correlations among the items which are used to measure his factor scales. Cattell also presents in the same article, as he has elsewhere, another explanation of how two items measuring the same factor can have a low correlation. Two items which have high positive loadings on factor I and loadings of $-.71$ and $+.71$, respectively, on a second factor can nevertheless have a 90 degree angle between their vectors. He does not give an example of two actual items which display this bizarre phenomenon, and it is presumed that he makes no claim that this is what causes all the low correlations among the items which are supposed to measure the same univocal factor in his factor scales.

Cattell makes a great point of the fact that a factor may emerge in an analysis because of a "swollen specific," (p. 757) a phenome-

non illustrated earlier under a different name by the reviewer (Comrey, 1961). Thus, there is agreement by Cattell and the reviewer that inclusion of highly similar items in a factor analysis may lead to the emergence of factors which are not of great importance, *per se*. One of the articles in this volume (p. 94) exhibits at least two factors of this kind, although Cattell interprets them as regular factors rather than as "swollen specifics." For example, the two most highly loaded items on factor I were, "42. Do you like to climb trees? (a) yes (b) no" and "70. Would you rather: (a) climb a tree, or (b) look at a book?" The third factor had the following two items as the most highly loaded variables: "26. Would you rather (a) go to school, (b) go on a long trip in the car" and "94. Would you rather (a) go to school, or (b) work at home?" Factor XV exhibited a similar phenomenon. These results demonstrate that even when one is aware of the dangers of "swollen specifics," it is not always easy to avoid them in practice, nor indeed apparently even to recognize them.

It is the reviewer's opinion that some of Cattell's methodological techniques are open to question. For example, it is stated that his factors are obtained by blind rotation to oblique simple structure in which correct factor positions are identified by finding the hyperplanes. In one article Cattell claims to demonstrate the necessity of oblique simple structure by showing how certain known dimensions are recoverable using this criterion, much as Thurstone did with his famous box problem. First of all, it should be pointed out that showing how oblique simple structure works in one case does not imply that it will work in every situation. It would be just as easy to construct an example in which oblique simple structure would *not* recover the known dimensions.

The reviewer also questions the value of blind rotation. This can only force the rotater to depend exclusively on the improvement of hyperplanes, or some other mechanical criterion, which may or may not give the proper location of a factor. The reviewer has shown (Comrey, 1959) that with a substantial number of factors and variables, it is possible to start with almost any variable as a trial vector, through using Thurstone's analytic rotation method, and to locate a good hyperplane for that variable by allowing the axes to go oblique. Further disquieting evidence has been found in unpublished studies by the reviewer. He used an analytic rotation program which he developed to search for good hyperplanes. Starting with original unrotated factor matrices, he found that factors were obtained with excellent hyperplanes, but that in many instances the factors were highly loaded by variables which had low or zero intercorrelations. Such factors were thus complex rather than univocal and proved to be generally uninterpretable. Varimax rotations with subsequent oblique adjustments for these same matrices pro-

duced much more reasonable results. It must be reiterated, however, that no procedure can be claimed to have universal application for location of correct factor positions merely because it proves to be successful in one or more instances.

In a recent article included in the volume, Cattell admits the existence of "false" or "geometer's" hyperplanes which have $k - 1$ points, where k is the number of factors. These hyperplanes, he asserts, must be distinguished from the "true" hyperplanes which have only about 70 percent of the points in them. The unwary investigator, he warns, may rotate to a "false" hyperplane instead of to a "true" one. In the course of blind rotations, the reviewer suspects that it may be very difficult indeed for even a wary investigator to recognize which is which. In any event, the number of points in the hyperplane for a given correctly-positioned factor will be determined by the particular combination of variables, number of factors, and perhaps other considerations. It will not necessarily be closer to 70 percent than to some other figure. Cattell actually allows a considerable divergence from this figure in his studies. The rotational solution is merely an interpretation of the data which should be based upon everything the investigator knows or can find out. There is no magical, mechanical procedure, such as rotation to oblique, simple structure with 70 percent of the points in the hyperplanes which provides a royal road to psychological truth. Cattell has not proved that his methods achieve this desideratum. His use of these procedures may be one reason why certain of his factors seem to be so difficult for at least some investigators to understand or to verify in their own investigations.

The fact that this review has meted out more criticism than praise should not be interpreted as an accurate index of the overall merit of this volume or the work which it samples. The articles reprinted in this book and Cattell's other contributions have earned for him a well-deserved international reputation. His original contributions will have an important influence on psychology for many years to come.

REFERENCES

- Comrey, A. L. "Comparison of Two Analytic Rotation Procedures." *Psychological Reports*, V (1959), 201-209.
- Comrey, A. L. "Factored Homogeneous Item Dimensions in Personality Research." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 417-431.
- Levonian, E. "A Statistical Analysis of the 16 Personality Factor Questionnaire." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, XXI (1961), 589-596.

ANDREW L. COMREY
University of California, Los Angeles

An Introduction to Psychological Statistics by Philip H. Dubois.
New York: Harper and Row, 1965. pp. xi + 530.

If one were seeking a single phrase by which to describe Dubois' new book, "extremely attractive" might come to mind fairly early. And the Phrase would be far from referring only to the delicate shades of turquoise, violet, and blue in which the dust jacket is printed. Within the covers of this volume is content of substance—content presented in a manner which is going to appeal to many.

There are eighteen chapters. As will become apparent, the scope—if not the depth—is ambitious. Prospective users may need to consider whether the claim of the book to be an introductory *text* fills their own expectations. This suggestion is intended as a *caveat* rather than as a criticism. It may well be that we have previously avoided too painstakingly the use of "advanced" mathematics in introductory texts. To disclose a bias, it is refreshing—if not exciting—to see matrix algebra appearing amongst the fare for an introductory course. But in case the mention of matrix algebra frightens a potential user, it must hastily be said that the bulk of the book requires little more algebra than has traditionally been demanded of elementary students in education and psychology.

The presentation of the topics follows closely the order established by Stevens' article, "Mathematics, Measurement, and Psychophysics" in the *Handbook of Experimental Psychology*. Dubois' acknowledged debt to this approach has resulted in a developmental sequence of subject matter which has both theoretical and pedagogical merit. Another merit is that the book has, throughout, steered a course well within the disciplinary boundaries of psychology and education. Accordingly, students should generally be familiar and comfortable with material used for discussion, examples, and exercises.

The first chapter gives an over-all view of the aims and uses of statistics in experimental and applied psychology. The two following chapters are based on description-by-enumerations or counting. By starting with categorical data, the author has minimized the need for complex mathematical symbolism early in the course. Certainly, the strategy permits a gradual introduction of unfamiliar notation. But one could perhaps ask whether the author (or the printer) has not created a potential danger in some usages of his symbolism in a desire to prevent "symbol shock." In these early chapters, we occasionally find mathematically doubtful expressions such as $(\Sigma f_e/f_o)$ for $\Sigma(f_e/f_o)$, and are told that "probability" . . . "operates." These lapses are in such strong contrast to the crispness and relative rigor of subsequent sections of the book that it is a pity they occur at all.

Chapter 2, then, outlines the nature of categorical data and discusses the contingency coefficient and the expectancy chart. Chapter

3 gives a very clear introduction to chi square. One of the real attractions of the book is the liberal use of illustrative examples—and they are used with special effect in these early sections of the text.

Chapter 4 shifts to ordinal data. Rather strangely, perhaps, the concept of making a frequency distribution is introduced for the first time. Other topics covered include percentiles, rho, and tau. Chapter 5 takes up the question of interval and ratio data and proceeds to derive and to illustrate the standard "averaging" techniques. For an introductory course, the matter is presented in as explicit a fashion as could ever be asked for—the only reservation one might have is the use of V_x for variance, hardly a transferable symbol.

Chapters 6 to 9 cover the standard parametric correlation indices including multiple correlation, partial correlation, the correlation ratio, and variants of Pearson r . These chapters which are all exceptionally well presented have commendably set out to show the interrelationships of the different indices. In many respects, the breadth of coverage is startling. There is a brief mention of suppressor variables; there are suggested applications for part correlation; there are exercises involving eta. In other words, there is basic material that will be interesting and challenging to the potential specialist.

Chapter 10, called "Forecasting Human Behavior" is an excellent summary and consolidation of the preceding four chapters, indicating practical applications and interpretations.

Chapters 11 to 13 introduce the normal curve and the inferential statistics dependent thereon. Again one is impressed with the variety of topics covered within these ninety pages; it might be, however, that the attempt at breadth is made at the expense of depth, even for a first course. The t distribution and F are treated in four pages—only one more page than for the Poisson distribution.

One of the most surprising aspects of the book is its relatively light emphasis on analysis of variance (ANOVA). Only simple one- and two-way ANOVA are attempted in Chapter 14, and these on a partitioning-of-variance model. What is attempted is done well, and there are some good exercises.

Chapter 15 summarizes briefly statistics used in test construction and interpretation. There is an apparent reluctance to state the conditions for parallel forms, but a considerable amount of ground is covered in the thirty pages allocated to the topic.

Chapter 16 is devoted to the use of matrices and determinants in elementary statistics. Chapter 17 goes on to apply the matrix notation to elementary factor analysis. In view of the widespread use of this technique, the inculcation of the ability to read data presented as factor scores is a desirable innovation. Both chapters

do no more than scratch the surface of their respective topics but appear to achieve the aim of enabling the student to read a factor analytic report. No serious attempt is made to introduce computational techniques, but a useful general discussion is given.

The final ten pages of text give a very brief outline of the more useful non-parametric tests, such as the run and sign tests, Mann-Whitney U , and the median test.

The Appendix consists of the tables referred to in the text and a glossary of symbols. The index is full and easy to use. Each chapter ends with a selection of key references for further reading.

The value of the book must be judged on its potential worth as an introductory text, since its depth is not great enough to place it in a reference class. On the whole, the prospective buyer would need to decide only whether the emphases in the book are isomorphic with his own. For clarity of exposition, for excellent illustrative examples, and for the availability of exercises, the author deserves praise. For some symbolic usage and a certain restlessness of content, one might feel a little uneasy. If one were thinking of using the book in his own courses, it might be worth considering whether one's teaching strategies were enhanced by, or fitted to, the text. Certainly the book makes ample provision for individual differences, which would free an instructor to place emphasis where he wishes. For the student, the book is eminently readable.

Dubois' book deserves a great deal of attention. It is to be hoped that it receives this well earned attention.

PETER A. TAYLOR
University of Illinois

Concepts of Statistical Inference by William C. Guenther. New York: McGraw-Hill Book Co., 1965. pp. xii + 353.

Guenther's recent volume joins the flow of new texts intended for use in a first course in statistics. In order to stand out from this ever-increasing flood, introductory texts are having to satisfy increasingly specific demands. Here is a fine book that may drown, unfortunately, in its very struggle to avoid a definite adherence to any one discipline that—in psychology or education, for example—might have won it many proponents.

Guenther is, of course, a statistician. He writes as a statistician, trying to communicate the ideas and language of statistical inference "without unduly slanting the examples and problems towards any field of application." As a mathematician, one has no hesitation in applauding the attempt; but as for its usefulness for teaching statistics in education and psychology—with all the inherent difficulties which *that* usually implies to both the student and to the instructor—one might have some doubts.

The major problem facing the potential user of this text in edu-

cation or psychology would be to decide whether or not his students were adequately prepared mathematically. The *operations* required of the student are far from complex—a working knowledge of high school algebra would be sufficient. However, the very early exposure in the book to a wealth of symbolism and vocabulary may overwhelm the mathematically unsophisticated student. The potentially frightening array is somewhat heightened by the particular developmental sequence adopted in the book wherein probability, expectancies, combinatorial analysis, and summation all appear within the first twenty pages, and not long after come exponentials, inequalities, and the multinomial theorem. Thus the utility of the book to the student would in a large degree depend upon the extent of his mathematical vocabulary. If the latter were impoverished, much of the *statistics* might pass by, hidden behind a semantic and symbolic barrier.

Lack of acquaintance with the particular vocabulary of a discipline should not, however, disguise the real value of this text. In education and psychology, the fact that too few students are immediately comfortable within a mathematical context may reflect inadequate demands for preparation for a course in statistics. Furthermore, it may well be that there should be a greater willingness to accept the need for a different kind of preparation for those students who want no more than a knowledge of statistics sufficient to help them read research reports, from those whose interests would take them far deeper. For this latter group, Guenther's text provides a sound introduction. For the former group, the rapid exposure to awesome symbolism, the lack of examples specifically attuned to education and psychology, and the rigor and tenor of the presentation, may leave them dizzily grasping for straws. The contention is, therefore, that this book should be evaluated in the light of its utility as a textbook for the mathematically more mature.

Having settled upon a preliminary framework within which the book's usefulness in education and psychology is to be judged, the reviewer next seeks to determine whether the author's objectives in writing the text have been achieved. Briefly, the objectives are:

1. to provide material for a one-semester introductory course in statistics
2. to introduce students to the language and philosophy of the statistician
3. to acquaint students with the kinds of problem that lend themselves to statistical solution
4. to enable the student to work through some standard statistical problems
5. to enable the student to read research reports and statistical summaries with understanding
6. to encourage, among some students, a further study of statistics

These far-reaching objectives are best considered one at a time. With respect to the first objective the general scope of the book is not much beyond what is usually regarded as suitable content for an introductory statistics course. The title of the book indicates the emphasis—on statistical inference—yet the descriptive statistics receive due treatment.

The first chapter presents the basic definitions and manipulations of probability, permutations and combinations, random variables, probability distributions, expected values, and continuous random variables. No one topic is treated to any great depth. In fact, it is the rapid transition from concept to concept that creates the potential hazard for the naive student. The approach tends to be one of explication by example. Exercises are provided for the student to work following each major point. (Answers are not always given to all the exercises, and the general difficulty of the exercises tends to be higher than the textual illustrations.) The over-all effect of this first chapter, then, is a rather disjoint series of statements separated by problems, the student's success with which is probably not sufficiently obvious to himself, owing to lack of feedback in the form of a complete set of solutions.

Chapter two, in contrast, provides a neat, concise exposition of the basic probability models: the binominal and multinominal models, the negative binominal and the uniform probability models, the hypergeometric, and Poisson distributions. As if all this is not impressive enough, the student is hastened through the normal distribution, and the concept of degrees of freedom with reference to t , F , and the chi-square distributions. For one who has already had some exposure to the content, there is considerable satisfaction in these basic models being juxtaposed to the advantage of similarities amongst them. Again, there is nothing too fearsome in the depth—indeed, the treatment of each distribution tends to be rather superficial—but the rapidity of movement from one model to another may be more than a little disheartening to a student facing the concepts for the first time. The continuity is much more obvious than in the first chapter, however, although the same expository format is followed. A high reliance on repeated use of graphical diagrams to illustrate the points should contribute to effective instruction.

Chapter three covers parameters, as distinct from statistics, and sampling distributions. The mean and median are disposed of in little over a page, and yet no real feeling of loss is engendered. Similarly, measures of variability are covered in but four pages. The chapter then proceeds with a discussion of the normal distribution, as well as of the sampling distributions of the sample mean and variance. The standard score, and its derivative, which are traditionally much used in education and psychology receive scant attention. There is only one illustrative example, and no student

exercises, using material relevant to either psychology or education.

Chapter four provides an introduction to the concept of a statistical hypothesis. The discussion is again very stimulating—provided that some prior acquaintance with the content was made available, as, for example, in a teaching session. There is an overhaste to discard one topic for the next; and although everything one could wish for in an introductory course is present, the highly condensed expository style would leave even the better students somewhat dazed.

The next two chapters are the high point of the text, and of themselves deserve special attention. In the fifth chapter the author discusses statistical inference for means and variances; in the sixth, inference for parameters of discrete random variables. A reader who has mastered all this before will realize the elegance of the presentation. The development is finely revealing and immensely satisfying. For once, the book does not appear to be in danger of losing the student; and since, after all, this is probably the main point of the volume, these two chapters must be counted heavily in its favor. Topics covered include: testing hypotheses concerning the mean when the variance is known and when it is unknown; power calculations; confidence intervals; inferences about variances; inferences about the binomial, Poisson, and multinomial parameters; and chi-square tests. Each chapter ends with an excellent summary table. Examples are good; exercises—again not specifically oriented to education and psychology—are also good, though they suffer from lack of sufficient numbers of answers given.

Chapter seven leads naturally into the analysis of variance (ANOVA) technique. Like the two prior chapters, this chapter is much more relaxed. ANOVA is approached from the sum-of-squares model, although brief mention is made of the linear model. Only one-way ANOVA is attempted; and the method of presentation of this basic model, which is clear and precise, leaves no doubt as to the logical processes involved.

The final chapter discusses regression and correlation in a twenty-page summary. The stress is on potential inferences. There is a tendency for regression (sic!) toward the saltatory style of writing of the early chapters. The content is there, even if it has to be pried from tightly phrased sentences.

The appendix, extending over about eighty pages, not only provides notes on the summation notation, and on the use of square root tables and hand calculation of the same, but also furnishes several standard tables, as well as a series of graphs. The latter cover some extremely interesting topics: the power function of the t test; required sample size for given confidence intervals; power curves for the testing of hypotheses concerning variances as well as for F ; and confidence belts for the correlation coefficient. All of

these graphs could be most valuable. The actual size of the graphs may prevent a high degree of accuracy in reading them, but their presence in printed form in this volume is useful. Answers to exercises, and a small, but adequate index, round out the book.

To return to the first objective—content for a one-semester course. Clearly there is plenty of content in these pages. The author, who himself acknowledges the possible difficulty in getting through all the material, suggests that in such event, chapters six through eight should be omitted. For teaching purposes, much of the coverage would have to be supplemented—not in terms of further content, but in terms of an expanded presentation. For those who enjoy teaching to supplement a text, rather than teaching the text, this book provides ample opportunity for a satisfying classroom experience.

The second objective—to introduce students to the language and philosophy of the statistician—has also been admirably met within the confines of the needs of the intended audience. The logic of statistical method is repeatedly, and clearly, illustrated. The statistical-mathematical vocabulary is rich—perhaps too rich.

The third objective, to acquaint students with the kinds of problem that are most suited to statistical solution, is an ambitious objective, and is, in fact, the source of a weakness of the book for use in psychology and education. The illustrative examples in the text are drawn from agriculture, economics, games, industry, and the abstract, as well as from the immediate fields of interest. The problems which are all well chosen and interesting, lead to useful insights. For those students and instructors who prefer to work solely within their discipline, however, the diversity will be distracting. It sometimes causes the point of the problem to be hidden behind an interesting facade. The reviewer polled a number of students as to their preference for the content of problems; and the response, even from very competent students, was unanimously in favor of problems within the education and psychology domain. The inference is that students may find the wide variety of problem content unpalatable.

The fourth, fifth, and sixth objectives, which would have to be evaluated against student performance, require empirical verification. One can say little more at this stage than that the book *appears* to be potentially capable of bringing a student to a level of competence where he can work through simple problems and can have sufficient vocabulary to read most research intelligently. How far it will encourage students to a further study of statistics is anyone's guess.

Here, then, is a book that would be a highly valuable addition to the bookshelf of any instructor of elementary statistics, in any discipline. It would also be a worthy addition to the library of many students who propose to advance in statistics or who appreci-

ate mathematical elegance. For both these categories of persons the value of the book in education and psychology should not be overlooked. The presentation is refreshingly concise, and the too few references which are given are really most stimulating.

For the bulk of students one is usually confronted with in either education or psychology, there is a certain inappropriateness in both style and illustrative content. To get maximum return from use of the book with the less capable student would, it seems, require more time and effort than most teachers and students can afford.

PETER A. TAYLOR
University of Illinois

Reduced Rank Models for Multiple Prediction by George R. Burket.
Psychometric Monograph Number 12. Richmond : William Byrd Press, 1964. Pp. xi + 66. \$4.00.

Psychological test construction has progressed to the point where predictive instruments that will relate to almost any criterion variable can be composed. It is often difficult to build a *single* predictor that shares enough of the characteristics of a criterion to correlate very highly with it, but it is relatively easy to sidestep this problem. One merely constructs a number of predictors, each sharing somewhat differently, but together quite comprehensively, in the several aspects of the criterion, and uses some linear combination of them to give a best estimate of the criterion. A highly complex criterion may require a large number of predictors to represent it adequately.

In the case where a complex criterion represents success on a job or in a professional field of endeavor, one may find that there are actually fewer persons entering the field who can be tested than there are potential predictors of success in the area. This is especially true if a given occupation requires a large amount of ability. It is in this type of situation that a conflict arises.

To make the most accurate predictions, one must retain as many of the relevant and non-overlapping predictors in a battery as possible. But in order to find stable weights, i.e., ones that will hold up well on cross validation, to apply to the predictors in a regression equation, one must have substantially more observations than predictors. This requirement has traditionally been referred to as overdetermination. One aspect of the problem to which Burket has applied himself is that of keeping as much relevant information in the predictor matrix as possible while at the same time reducing the number of terms in the prediction equation enough to obtain stable sets of weights when sample size is small.

It is well known that in the extreme case where the number of observations is equal to or less than the number of predictors, that the least-squares weights for an arbitrary subset of predictors

(or more generally, any lower rank approximation to the matrix of predictor values) will give better predictions in a new sample than will weights for the entire set. As the sample size becomes larger with respect to the number of predictors, Burket argues, one must arrive first at a point where particular ranks and particular rank reduction methods are preferable to an arbitrary subset of predictors and finally at the place when the entire set of predictors gives better prediction than any lower rank approximation.

On this basis, the solution Burket presents is a reduced rank solution, and the first problem he confronts is determining which of the many methods of rank reduction is most efficient for this purpose. He next examines several approaches to specifying the optimal rank of solution and to establishing an index of the accuracy of prediction to be expected in new samples.

In attacking these problems, Burket chooses generally to use the regression approach rather than the theoretically more edifying correlational treatment because of the former's much greater simplicity and "... with the hope that the practical differences between conclusions drawn from the two models will be negligible (p. 2)."

Five reduced rank models are empirically tested in the study. Predictor selection methods are most often thought of in connection with the situation where an optimal set of predictors of a given size (usually determined by cost) is the goal. Nevertheless, as is pointed out, when predictor selection aims to increase accuracy of prediction by increasing the available degrees of freedom, it becomes a special case of rank reduction. Burket includes both the predictor accretion and the predictor elimination methods in his study. The other reduced rank models examined are: (a) principal-axes factors giving the highest multiple correlation with the criterion; (b) the method of approximating the correlation matrix by the largest principal-axes; (c) the method of approximating the inverse of the correlation matrix by the smallest principal-axes.

Despite several elaborate and quite elegant mathematical demonstrations, the selection of the reduced-rank models to be empirically evaluated within the regression theory framework is based largely on the author's judgment. Burket does show conclusively that, assuming the factors are to be chosen independently of criterion observations, the largest principal-axes factors will give the greatest expected accuracy of prediction in terms of total squared errors of prediction. This point is emphasized somewhat more than it deserves in order to provide a theoretical backdrop for the resounding failure of Guttman's method of approximating the inverse (using the smallest principal-axes factors) in the empirical demonstration that follows.

The empirical comparisons of the five techniques and indeed, of the entire procedure employed, are quite an impressive achievement.

Twenty-nine predictors and five separate criteria are used. Fifteen of the predictors are those composing the University of Washington Entrance Battery; the remaining 14 are from the Edwards Personal Preference Schedule. The criteria are grade point averages in Mathematics, English Composition, Chemistry, Psychology, and All-University areas. The observations are on 973 students for whom measurements on all predictors and on at least one criterion are available. The smallest criterion group is represented by over 500 cases.

To make the comparisons, the following procedure is replicated for each criterion and for samples of size 255, 210, 165, 120, 75, and 30 cases. A random sample is drawn from the statistical population of cases for whom measurements on a particular criterion exist. Regression weights are computed for each reduced-rank method and for each rank from 1 to 28 and for the full rank-29 case. Then a new random sample which always consists of 252 cases is drawn from the remaining cases, and the weights computed on the original sample are applied to it. The correlation between the predicted and the observed criterion in the new sample is used as a measure of accuracy of prediction.

For the All-University criterion an additional set of replications is carried out. These omit Guttman's method and employ, in addition to samples of the original sizes, samples of size 435, 390, 345, and 300. Two new random samples of size 252 are then drawn, and measures of accuracy of prediction computed. In addition to the correlation between predicted and observed criterion measures in the new samples, total squared errors of prediction are also computed. Since predictor and criterion variables are normalized in every case, means and sums of squares are equated and total squared error of prediction are therefore directly comparable.

The results of the study are generally clear-cut. For Guttman's method, the correlations between predicted and observed criterion values in the new sample are substantially lower than are those for any other method including full rank weights for every rank, criterion, and sample size. In 26 of 30 samples, the method of largest principal-axes most often gives the highest correlations. The exceptions are for samples of size 210 and 255. There is some tendency for this method to be relatively superior for the Mathematics and Psychology criteria than for English Composition and Chemistry. For the ten additional samples using the All-University criterion, total squared errors of prediction for each of two cross-validation samples are computed as well as correlations. Generally, the lower rank errors of prediction based on small samples and using largest principal-axes compare well with those for full rank and large samples. For the smaller original sample sizes, the largest principal-axes method is definitely superior to other methods, and lower rank

approximations are superior to higher. Even for the largest original sample sizes, the method of largest principal-axes seems to be preferable to full rank solutions.

Burket concludes that the method of largest principal-axes could be used to considerable advantage when, for a given sample size, one is interested in the greatest accuracy of prediction obtainable, or in the case when, for a given accuracy of prediction, one wishes to use the smallest original sample size. But to obtain the coefficients for any reduced-rank equation, the particular rank must of course be calculated first. Burket derives formulas to estimate both correlations between predicted and observed criterion values and total squared errors of prediction in order to determine their usefulness for the purpose of pinpointing an optimal rank. Unfortunately, while both estimates are fairly accurate, because of their respective large standard errors, neither is a satisfactory basis for choosing a particular rank. Given a certain amount of judgment, the index of squared errors of prediction can be helpful, but it is a far from adequate solution to the problem. Burket theorizes that the small success of his statistics may be due to the fact that the regression approach, which considers only the criterion variable random, does not allow the sampling variation of the factor loadings to be taken into account. He speculates that an analysis of prediction problems based on a model such as the multivariate normal correlation model (which assumes the predictor variables to be random) might lead to much more successful estimates of accuracy of prediction.

Burket's study must be characterized as ambitious and well-executed. Even the details have a glossy sheen. The presentation, the mathematical development, and the results are all clear and even boast a certain elegance of simplicity. The conclusions are well drawn and well supported. It is nevertheless true that they will surprise very few people who are working in the area. The value of Burket's work lies not in its being a major advance in the field of multivariate prediction, but in the fact that it consolidates a great deal of previous work, provides a sound basis for choosing among a number of alternative procedures, and points with certainty toward the most fruitful avenues of future research.

JAMES A. WALSH
Department of Psychology
Iowa State University

Introduction to Educational Measurement (Second Edition) by Victor H. Noll. Boston: Houghton-Mifflin Company, 1965. Pp. xviii + 509.

Eight years elapsed between the first and second editions of this text. Much had happened in the field of measurement. The present

revision of this popular text was prompted by the need to incorporate recent developments.

The number of pages has been increased by 15 percent over the first edition. A third appendix on extracting square root by the long-hand method has been added. Addition of the now-famous normal curve from *Test Service Bulletin* No. 48 of The Psychological Corporation of p. 52 is a worthwhile feature. The distinction between "ratio IQ" and "deviation IQ" on pages 60-63 is also a valuable addition. With two exceptions, chapter titles are identical with those in the first edition. The chapter on intelligence and aptitudes in the first edition was split into two chapters in the second. Section titles are nearly identical with those in the first edition, while section content has been improved in clarity and scope.

The present reviewer also reviewed the first edition for this journal (*Educational and Psychological Measurement*, XVIII (1958), 533-538. Because of the great similarity of the first and second editions, favorable comments from the first review were not repeated in the present review. However, an evaluation has been made of the extent to which the author has implemented critical suggestions for improvement from the previous review. Two inaccuracies, one in spelling and one in a historical date were corrected in the new edition. Three matters of debatable symbolism and terminology for certain concepts in variability and standard scores have been clarified satisfactorily. A suggestion of greater relative coverage of teacher-made tests as compared with standardized tests has been realized in a more balanced coverage. The suggestion of the need for an instructor's manual was followed by the appearance of one shortly after publication of the first edition, and a revised manual for the second edition has been promised by the publisher for fall. Still another suggestion regarding the relative coverage of sources of information about standardized tests as compared with specific detail about selected standardized tests shows slight improvement. The only critical suggestion which seems not to have been heeded was in reference to a sentence in the section entitled, "Meaning of Correlation." In the second edition it is on page 54 and is identical with the corresponding sentence in the first. It refers to perfect positive correlation and reads as follows: "This exists when a change in one trait is always accompanied by a *commensurate* change in the *same direction* in the other trait." It still appears to be potentially misleading for the beginning student of measurement.

Despite minor criticisms, this book is highly recommended for consideration as a textbook in the introductory measurement course for undergraduates in which either or both of the elementary and secondary teaching curricula are included. If the revised *Instructor's Manual* is as good as the first, it will also weigh in the decision

of whether to adopt this text. The first manual was as good as any the reviewer has seen.

SAMUEL T. MAYO
Loyola University (Chicago)

Interpretation of Test Results by Kenneth F. McLaughlin. Washington, D. C.: U. S. Department of Health, Education, and Welfare, Office of Education, 1964. Pp. vi + 63.

According to the foreword by Arthur L. Harris, Associate Commissioner, "... This bulletin attempts to explain the use and limitation of regularly administered tests, so as to enable administrators, counselors and teachers to interpret better their meaning to parents and students." Apart from occasional forays into esoteric concepts, the general tone of presentation appears to be adjusted to the level of the testing "layman" who may be overwhelmed by more complete measurement texts. Although many recent criticisms of testing have been directed toward interpretational faults, the guilty parties are so often unable or unwilling to find the time to inform themselves adequately of the implications of their dubious practices. If only the present little volume could be read by those who need it most, much could be done to dispel the stigma of inadequacy that has pervaded the field of measurement in the past few years.

By beginning his introduction on a negative note, McLaughlin is parroting an apologetic trend that seems to be creeping into many similar booklets. Although proper cautionary advice is pertinent in its proper place, it hardly seems necessary to begin with a pull of the rug which virtually undermines any confidence that one may have in the testing profession. Surely there are many positive advantages of careful testing and interpretation that could be used to motivate the reader to improve his techniques and to restore or build a dynamic faith in the merits of his chosen field of endeavor.

The preliminary chapters touch lightly on topics such as test construction, various types of tests, scoring, and score accuracy. Then follows a much more detailed discussion on ways and means of analyzing class performance and interpreting results to students and parents. Of particular value is a very excellent list of selected references covering almost all aspects of elementary measurement theory and practice.

It is relatively easy to find inadequacies in a book of this nature, because the author was obviously limited by space and also by the need to keep within the bounds of the layman's understanding. In attempting to be brief—such as in his discussion on intelligence, mental ability, or scholastic aptitude tests—he succeeded in providing a clear picture of the outmoded ratio IQ, but very little else. Inasmuch as IQ interpretations are often a major nemesis of

counselors and teachers, it would seem appropriate to do more justice to this important concept.

But a more glaring inadequacy was found in the handling of percentiles and percentile ranks. This was revealed in the application of the standard error of measurement to percentile bands, in the discussion on scattergrams, and in the section on profile interpretations. First of all, there seemed to be a need for a more adequate understanding of the terms themselves. A percentile is a *score* having a specified percentile rank, whereas the percentile rank is the *status* or rank position of that score in the total distribution. That is, the 75th percentile is the upper limit of the range of scores attained by at least 75 percent of the students. Such a score has a percentile rank of 75, but this does not mean that the percentile is 75.

To complicate the issue still further, the author defines a percentile as the percentage of individuals in the norm group who made scores *below* any given score. "For example, a student at the 75th percentile scored better than 75 percent of the students in the norming population." A more accurate definition would have been that the 75th percentile is the upper limit of the scores attained by at least 75 percent of the group. Although such a score has a percentile rank of 75, this does not mean that 75 percent of the cases are below the 75th percentile or that a student with this percentile has a score *better* than 75 percent of the group. The next lowest score in the distribution may have a percentile rank of only 70, thus encompassing only 70 percent of the group. In such a situation, it should not be said that the student has scored better than 75 percent of the group, for in actual fact he is superior to only 70 percent.

In the bivariate scattergram given on page 23, there is some confusion as to whether scores, percentiles, or percentile ranks are intended. Furthermore, in a chart such as the one suggested, and also in the presentation of profile data as discussed in a later chapter, certain assumptions must be made regarding the nature of the groups on which the percentiles (or percentile ranks) have been determined. It is erroneous to suggest that percentile ranks based on one group may be plotted against percentile ranks based on an entirely different group to provide a means of interpreting achievement in relation to ability. No doubt the author is aware of this type of error in interpretation, but there is no word of caution to guide the unwary novice away from the obvious pitfall.

A second problem with regard to scattergram analysis is introduced through the suggestion that a diagonal line drawn from corner to corner will provide a means of making expectancy-type interpretations, especially if approximate error bands are plotted on each side of the diagonal. This kind of graphical presentation is

often one of the primary reasons for much error associated with the identification of underachievers and overachievers. A straight diagonal line is not a logical expectation where regression is operating, and it may be wrong in many testing situations to use the diagonal as a means of determining that such and such a score on an achievement test is "consistent" with ability.

Perhaps some of these semantic problems are the result of personal differences between the author's concepts and those of the reviewer, but they occupy such a large portion of the book and are so important in proper test interpretation that their presence detracts considerably from the many fine aspects of useful information that are included. Although the booklet may prove to be a great help to teachers and counselors who have already been exposed to many of the hazards of measurement, there may be some question in the desirability of allowing anyone to gain the impression that herein lies all one needs to know in order to function as a reasonably informed interpreter of tests. From a professional point of view the booklet has much to offer and much to be desired. Used with discretion it should certainly contribute to a better understanding of the merits and possibilities of educational measurement.

NORMAN C. MABERLY
Test Department
Harcourt, Brace & World, Inc.

Theory and Problems of Statistics by Murray R. Spiegel. New York: Schaum, 1961. Pp v + 359. \$3.50

Theory and Problems of Statistics is one of a group of volumes issued by Schaum Publishing Company in its *Outline* series. The purposes of the book are "to present an introduction to the general statistical principles which will be found useful to all individuals regardless of their fields of specialization" and to serve as a reference book for workers engaged in applications of statistics. The only mathematical background assumed is arithmetic and algebra. A major emphasis is placed on solving problems, and 875 problems "completely solved in detail" are presented. This volume is intended for an introductory course in statistics, primarily as a supplementary text, but also perhaps as the basic textbook.

Since this book is intended to be useful to students and workers in a variety of disciplines, the coverage is both different from and broader than is typical in statistics books prepared for psychologists. The first five chapters deal with variables and graphs, frequency distributions, measures of central tendency, measures of dispersion, and moments. Next follows a group of six chapters which attempted to present in an elementary form the concepts underlying the use of statistics. These include probability theory; binomial, normal, and Poisson distributions; sampling theory; statistical estimation theory;

statistical decision theory, including tests of hypotheses; and small sampling theory (including a treatment of "Student's" t). Two chapters concerning respectively the chi-square test and curve fitting are followed by chapters on correlation theory and multiple and partial correlation. The book concludes with a chapter on analysis of time series and one on index numbers. Eight tables necessary to solving many of the problems and to using the various procedures are presented in an appendix. A useful index has been provided.

As is to be expected in an "outline" book, the treatment, which is very terse, would require much expansion by the teacher if this volume were used as the primary text in a statistics course. The treatment is generally lucid, however, and appears to be accurate. No errors other than typographical were observed. The broad coverage of topics could be used both to give psychology students a "feel" for how other disciplines use statistics and as a preparation for inter-discipline collaboration. The reviewer, who has found this book a useful reference, would like to use it (as a supplementary text) in teaching a statistics course. In short, within the limitations imposed by its purposes, *Theory and Problems of Statistics* is an excellent textbook.

JAMES M. RICHARDS, JR.
American College Testing Program
Iowa City, Iowa

Personnel Interviewing: Theory and Practice by Felix M. Lopez, Jr. New York: McGraw-Hill Book Company, 1965. Pp. ix + 326. \$8.95.

This is a comprehensive text on interviewing procedures intended for a wide variety of readers who perform interviewing functions, but presumably it is also meant to be used as a basic text in personnel management courses. The book is well written and well documented. There are 111 references scattered throughout the 326 pages of text. Most of these are up-to-date citations drawn from the professional literature, mainly within the discipline of psychology. The author's approach to interviewing is eclectic; his style of writing is more narrative than esoteric; and his purpose for writing the book is to restore the dignity and value of the interview as an important tool of personnel management. All of these combine to render the book interesting, informative, and useful.

The book is divided into four parts. In Part I the author presents background data to the study of the interview process which he defines as a communication medium; he also defines other terms to be used later in the text, and discusses some basic principles of psychology, communications system, and role behavior. Part II consists of discussions on the effective use of the information-exchange interview in personnel situations which warrant it in

occupational settings. In Part III the author discusses, with illustrations, the application of the problem-solving interview as an effective tool of personnel management. Part IV is devoted largely to a discussion of the decision-making interview with specific reference to the screening and selection of job applicants.

Although the comprehensiveness of the book has much to recommend it as a basic guide and reference book for interviewers or interviewers-to-be, this reviewer has some misgivings about the scope attempted by the author. As the title implies, the book is a treatise on the theory and practice of personnel interviewing. But it is much more than that! It is a potpourri of exposition on such topics as motivation systems, sensation and perception, frustration and aggression, self-concept theory, information theory, personality development and assessment, reactions to conflict, formation and measurement of attitudes, psychological testing, and statistical analysis of test results. Through his expositions on the foregoing subjects, the author nearly runs the gamut of the "miniature systems" of psychology, instructing the reader in the fields of social, clinical, experimental, industrial, differential, and counseling psychology.

Although it is undoubtedly true that the art of interviewing is substantially dependent upon the knowledge and application of psychological principles, this reviewer questions the advisability of attempting to impart such knowledge to the reader by including these topics among those related to the major purposes of the book. There is no adequate substitute for formal prerequisite instruction. For the reader who possesses a solid foundation in psychology, a review of these fundamental principles is unnecessary. For the psychologically naive reader, the superficial treatment of these topics through the brief dispersment of them throughout the text is not likely to improve the level of sophistication in psychology of the reader. In fact, misinformation and misinterpretation may result. In addition, although the author generally does a commendable job in his explanations and definitions of psychological principles and phenomena, there are some instances in which the uninformed reader might be seriously misled. For example, on the inside cover, the publisher (with the author's permission, it is assumed) tells the reader "Applying proven psychological principles to personnel management practices, the author shows you how to develop and enhance your skills in conducting interviews. . . ." Most psychologists would claim that few, if any, psychological principles have ever been proved! On page 60, the author tells the reader that the non-directive counseling and client-centered counseling of Rogers are the same, which they are not. On page 72, in distinguishing among factual data, logical data, and psychological data, the author includes among the latter only those data, such as ones relating to attitudes and values, which refer solely to the non-cognitive attri-

butes of behavior. He thereby leads the psychologically naive reader to an erroneous conclusion regarding the nature and scope of "psychological" data. Further, on page 221, where he discusses correlational analysis, he gives the reader the impression that the psychologist is satisfied with basing his predictions of a worker's success on-the-job on a single predictor variable—a condition which certainly does not prevail even among the least competent of psychologists!

In summary, this reviewer feels that Dr. Lopez would have made a significantly better contribution had he omitted his well-intentioned, but questionable, attempt to instruct his reader in general psychology. The prerequisite requirement of a minimum level of sophistication in psychology would most certainly have restricted the scope of the readers. However, it might have forced those readers who do not possess the understanding of fundamental principles of psychology, to gain it through the equivalent exposure to an undergraduate course in general psychology and thereby to profit more from the excellent treatment of the problems and principles of personnel interviewing, which after all, constitutes the author's objective.

PETER F. MERENDA
University of Rhode Island

Conflict and Creativity by Seymour M. Farber and Roger H. L. Wilson (Editors). New York: McGraw-Hill Book Co., 1963. Pp. xvi + 360. \$2.95 (paperback), \$6.50 (hard cover).

This volume is composed of the proceedings of the second University of California Medical Center symposium on "Control of the Mind." A reviewer of a symposium labors under considerable difficulties, since he can scarcely avoid some variation of "the content and quality of contributions." This cliché certainly applies to the present volume with a vengeance. Indeed, the mind boggles at just how heterogeneous this book is. It reads something like a random selection of papers from an APA convention. The content varies from detailed, technical explanations of experiments on the effect of drugs on behavior to William H. Whyte, Jr.'s discussion of difficulties in getting public and legislative action to preserve open spaces. The writing varies from turgid to entertaining, with, unfortunately, more of the former than of the latter. The interest value and quality of the papers are equally varied. Surely there must be some limit to what people who organize meetings and who publish books will try to foist on the public in the way of papers whose only connection with each other is that they were given at the same meeting or are between the same two covers. Whatever that limit is this book is pushing it. Moreover, the title is more an attempt to capitalize on current intellectual fads than an ac-

curate description of the contents. Such titling is another deplorable tendency of publishers of symposium proceedings.

The book is organized into six sets of three papers each, which correspond to individual sessions of the symposium. Following each set, this book transcribes the formal but spontaneous panel discussions which followed each session. As might be expected, some of the best ideas came out in these discussions, but they are difficult to find in the crowd of not-so-good ideas. The book concludes with an address by M. Hervé Alphan (Ambassador of France to the United States).

The first session was titled "Individual Potentialities." The first paper in this section by Benjamin Pasamanick on "Determinants of Intelligence" may well be the best and most important paper in this volume. Dr. Pasamanick, who points out that congenital conditions are not necessarily hereditary, describes his research on the relationship between poor prenatal care of mothers and intellectual deficit in children. He then states that the level of prenatal care of Negro mothers in the United States is certain to produce much intellectual deficit in their children. In addition to being yet another nail in the coffin of theories of inherent differences between races, this presentation has most important social implications. Since the intellectual deficit resulting from poor maternal care appears *irreversible*, corrective programs among the underprivileged, especially Negroes, at the pre-school and later levels may be important and commendable, but they are not enough. What is needed is a just society in which high standards of medical care and nutrition are provided to all. Nor should this just society be restricted to the United States. As Pasamanick points out, inhabitants of less advantaged nations must suffer from these factors to an enormous degree. In the second paper in this session, Gardner Murphy in the space of 14 pages covers the waterfront on "Determinants of Personality" from genes to national character. S. Rains Wallace concludes this section with an amusing and excellent chapter on problems (including ethical) of predicting performance on the job. He points out that so called "selection" procedures are actually "rejection" procedures.

In the second session titled "Clinical Measurement of Mood," Vincent Nowlis discusses the definition and measurement of mood. Keith F. Killian, Jr. shows that the effect of drugs on central response to peripheral stimulation depends in part on the "meaning of the stimulus," and Henry K. Beecher has a paper on Voodoo death, the placebo effect, and related phenomena.

The third session is closely related to the second, and is titled "The Evaluation of Drug Action." Leo E. Hollister writes of the difficulties in evaluating effects of psychotherapeutic drugs on mental hospital patients; P. B. Drews portrays the use of operant condition-

ing of animals (especially pigeons) in measuring drug effects; and Harris Isbell narrates the historical development of attitudes toward drug addiction in the United States.

Both the fourth and fifth sessions were titled "Conformity and Diversity." In the fourth session, Harley Cantril (as might be expected) describes life as a "transaction" between the individual and society; Rollo May writes of the meaning of freedom and determinism; and Richard S. Crutchfield discusses the dynamics of conformity and independence. Dr. Crutchfield actually presents empirical evidence. In the fifth session David G. Mandelbaum discusses the cultural context of conformity and diversity; Dr. C. H. Waddington points out that in the human species there is social as well as biological evolution and argues that human ethical feelings are an integral and probably essential part of the mechanism of social evolution; and Carl Rogers describes psychotherapy in terms of "learning to be free."

The sixth session, titled "Creative Expression," includes a presentation by the violinist Joseph Szigetti on the demands that serious music makes on both performer and audience, William H. Whyte, Jr.'s exposition on his efforts to preserve open spaces, and a paper by Flavio de R. Carvalho called "Notes for the Reconstruction of a Lost World: The Age of Hunger." The reviewer is still uncertain just what this last paper is about.

The final address by M. Alphonse is a rather elementary consideration of diplomatic negotiations in the resolution of differences between nations.

Like any heterogeneous set of papers, this book could be used as a book of readings in some undergraduate courses. The reviewer knows of one case in which it was employed effectively in a course on "Creativity and Man" taught collaboratively by a psychologist and an English professor. Moreover, for some readers one or two individual papers may be worth the purchase price. In general, however, the reviewer recommends against buying *Conflict and Creativity*.

JAMES M. RICHARDS, JR.
American College Testing Program
Iowa City, Iowa

The Counselor and Society: A Cultural Approach by Lawrence H. Stewart and Charles F. Warnath. Boston: Houghton Mifflin, 1965. Pp. xiii + 400. \$6.95.

A counselor is his brother's keeper: he not only is to "help clients to evaluate themselves and their environment and to effect an adjustment between their self-perceptions and environmental realities," but also actively seeks to "change the social conditions that hinder youth's struggle to attain a clear-cut sense of iden-

tity" (p. xi). This statement is the cornerstone on which the book is based. The counselor as a social engineer should "expose children to a variety of value systems," "confront youth with the potential consequence of their decisions," and "deliberately provide youth with experiences which would enable them to become purposive individuals . . ." (p. 44). These and other functions are "not necessarily incompatible with individual freedom as long as the ultimate decisions are left to the students" (p. 44). The student, like a dirty shirt that has been washed and wrung out, is left pinned to a line. Each one is free to follow his natural inclinations. It is hoped that the individual unlike the shirt does not dry up in exercising his right to freedom in ultimate decisions.

The authors cite three sources that affected the writing of the book. First, from Byrne (1963) they adopted his definition of counseling that includes ". . . helping individuals to obtain and maintain self-awareness, confront threats to being or becoming, and bring full fruition to their unique potential within ethical limits imposed by society." Second, from Tiedeman and Field (1962) the authors made use of the notion that "guidance is a process of helping individuals develop purposive behavior." Inherent in this concept is the idea of Festinger's cognitive dissonance. (It is unfortunate that the Tiedeman and Field article is not properly cited. For this reference the citation has been transposed between Chapters 2 and 3.) From Super (1951) they have used his idea "that vocational guidance is a matter of implementing a self concept." Third, using these sources, Stewart and Warnath formulate two basic functions of guidance: (a) help the student formulate a sense of identity; (b) help to provide a measure of reality so that progress towards goals can be provided. In order to reach these goals, the authors suggest three strategies: remedial, preventive, and promotional. Although the plan developed by the authors has some degree of coherence, they continually damage their case by insisting that individuals in this guidance setting have freedom. They do not acknowledge the fact that the manipulation, maneuvering, and mastering by the counselor restricts the flexibility of movement of the individual. What the counselor is trying to instill is a sense of duty to society at the expense of individual freedom. Hence, freedom, especially in the existential sense, can never exist in the proposed guidance setting.

Stewart and Warnath are not interested in pupil personnel services. They propose a concept of guidance services, although it is not so explicit as in some of the current textbooks on the subject. They do such topics as information, appraisal, cumulative records, counseling, group procedures, and research. Chapters 3 and 4 provide the psychological, sociological, and economic orientation to the rest of the text. Chapter 3 reviews the relationship between the values in-

herent in the guidance process and that of the school. It also discusses the relationship between the counselor's value system and that of the client. In addition three models of behavior—cultural, developmental, and end goal—are reviewed. Chapter 4 examines the problem of identity: How does a child become an adult? This problem includes such topics as perception, role models, school atmosphere, and socialization.

Chapters 3 and 4 should stimulate the individual counselor to look for background material in subject areas outside of education. The need for a broad background becomes apparent when a counselor tries to execute the two basic functions of guidance: identity and reality. It is unfortunate that the book has only a few examples on how to carry out the ideas contained within it. This shortcoming is not an oversight, for the authors state that "the book does not present a systematic explication of the operational procedures for implementing the guidance goals within the educational setting" (p. xii). They suggest that since there are many ways of achieving the goals of guidance, each counselor should determine his methods of reaching them. Ironically, it is appropriate for a counselor to be a social engineer with respect to his own students, but unfortunate for textbook writers to point the way.

This failure to take a stand is a basic weakness of the book. The authors raise many interesting questions, but fail to answer many of them because they say the answers are too complex. They should have pointed out the continuum of possible answers and stated where they stand. Many counselors who lack the sophistication to develop unique plans must look towards the "expert" for specific ideas.

Stewart and Warnath state that "counseling is personalized learning" (p. 256). It is a rational process in that "each case becomes an experimental venture: the posing of hypotheses, their testing, and the acceptance or rejection of the hypotheses" (p. 261). Since the primary goal of counseling is the "development of reasoning ability," then "every genuine counseling situation contains the same beginning and end no matter what happens in between" (p. 260). In other words *affect* tends to be of only slight importance in counseling. How an individual feels about matters seems to be relatively unimportant as long as the problem is rationally solved. Although the authors talk about not slanting material or accepting decisions of a client, it is implicit that rationally sound decisions should be reached. In the case of Sam (pp. 166-172), the authors in appraising Sam's decision to continue with a mathematics course, picture a squirming counselor who feels uncomfortable with Sam's conclusion. Should he interfere with it by informing Sam's parents? "As a counselor, probably he should have made this effort" (p. 171). Again the general statements on individual freedom, avoidance of

prejudice, antideterminism, and values are mere empty words. "I am my brother's keeper" is translated into meaning "you can do anything you want as long as it is done my way."

The reviewer's bias towards the book stems from the fact that Stewart and Warnath are inconsistent in the development of their theme of a cultural approach to guidance. The argument centers on the following dichotomy: Is the counselor an agent of the client? Is the counselor an agent of society? Since they believe in the latter point of view, it is rather difficult to talk about complete individual freedom. By candidly admitting that individual freedom is structured, they could have strengthened the book in many areas. After all, they do admit that the guidance process which they advocate is one of social engineering. Limiting individual freedom is a natural consequence of the process.

With the above remarks in mind, the book does have many positive highlights. The concluding chapters, "Staff Relations," "Ethical Considerations for Guidance," and "Professional Status of the School Counselor" have merit. The discussion of the rationale of the book makes for interesting reading.

Those who are interested in reorganizing their guidance program with an eye on the "Great Society" should read this book. The book will make them aware of some of the difficulties that they can encounter in making a culturally bound guidance program.

REFERENCES

1. Byrne, R. H. *The School Counselor*. Boston: Houghton Mifflin Co., 1963.
2. Super, D. E. "Vocational Adjustment: Implementing A Self Concept." *Occupations*, XXX (1951), 88-92.
3. Tiedeman, D. V. and Field, F. L. "Guidance: The Science of Purposeful Action Applied to Education," *Harvard Educational Review*, XXXII (1962), 483-501.

HENRY KACZKOWSKI
University of Illinois

Guidance Services: An Introduction by Carroll H. Miller. New York: Harper and Row, 1965. Pp. xi + 418. \$6.00.

The book is written in the spirit that the concept of "guidance" has grown to the point where it can rely on its own fund of knowledge rather than continue to lean for support on other subject fields. This point of view is not to imply that the guidance process should be divorced from the total educative scheme but that it has a separate identity of its own. As a consequence, Carroll Miller emphasizes "guidance services"—specific activities that are utilized to help the individual. What should be the direction of this help? Ac-

according to Miller, it should "assist students with their individual developmental needs."

Although "guidance" has developed into a specialized body of knowledge and skills, operationally it cannot be reduced to a set of definitions and rules. It has to operate within the context of the school setting. This operational orientation means that guidance is affected indirectly by the philosophy of the school and directly by the curricular emphasis. Since the guidance program must interact with these forces, Miller feels that it is impossible to state categorically that a guidance program must operate in a specific manner. This position by Miller is somewhat atypical, for the current trend in textbooks is to develop a systematic program of guidance within a single point of view. He also emphasizes the idea that "guidance services" must be developed within the context of "pupil personnel services."

Since a program of "guidance services" operates within a multi-dimensional framework (school, community, culture, etc.), some focal point is needed in order for the plan of operation to have some coherence. Miller utilizes the "self concept" for this end. Of the numerous themes that emerge, two tend to pervade the book. First, is the general idea that "planning" (educational or vocational) should receive a high degree of priority in the functioning of the "guidance services." Miller believes that "planning" is needed in order for a mature self to develop. He does not imply that guidance should be concerned only with prediction but that "plans" pave the way for stability and socialization of the individual. This notion gives rise to the second pervading theme: the school in general and "guidance services" in particular cannot and should not be solely responsible for helping the individual meet his developmental needs." Every facet of society should be mobilized to assist the individual. For example, in the area of vocational development, Miller feels that the high school can play only a minimal role in it because the student has just begun the "exploratory phase" when he leaves the school. He believes that agencies within the community should be set up to assist individuals with vocational developmental problems.

The ideas of "planning" and "total community involvement" will probably not be accepted by all who read the book. The latter concept represents a threat to some counselors because as "specialists" they will tend to lose status when too many "outsiders" are permitted to be guidance workers. The concept of "planning" restricts the scope of Miller's view on the topics of elementary school guidance and counseling. To a certain degree Miller finds it difficult for the counselor to deal with "personal problems." Educational or vocational matters can be readily discussed within the framework of "planning." However, some counselors believe that all counseling is

"personal;" therefore, "planning" plays only a minor role in their method of operation. Since the child on the elementary level lacks maturity, many of the activities outlined by Miller for this instructional level tend to be more "manipulative" than "planning." Miller does not imply that "planning" is done only by the counselor or that it is equivalent to "controlling," but that it is a societal demand. Nevertheless, the concept can bring about many interesting counter-arguments.

The book contains thirteen chapters. The first three chapters examine the various foundations of guidance. The chapter entitled "Guidance, Curriculum, and Learning" has an excellent review of the interrelationships between these three facets of education. The next four chapters deal with the elements of the guidance services. Miller does not follow the stereotyped classification in reviewing the activities contained in these services. In the chapter on "Understanding the Individual," Miller examines a large variety of techniques that are used by guidance workers. For the most part, the techniques are reviewed in a shallow manner. This lack of intensive treatment is atypical, for Miller tends to handle most issues in depth. The next three chapters are concerned with developmental matters. It is understandable that "vocational development" can be a separate chapter. But why does the author devote chapters to "Underachiever and Dropout" and "Superior Student?" Where are the chapters on the "Culturally Deprived Students" and perhaps more important on "The Normal Student?" These chapter headings point to the characteristic weakness of the service approach to guidance. It is primarily concerned with "problems" and with their identification rather than with students. The last three chapters are a potpourri: "Pupil Personnel Services," "Guidance Services in the Small School," and "Research in Guidance." The second listed chapter heading is a welcomed entry, for the operational concerns of a small school are somewhat different from those of a large school district.

The book is written for those who want a broad description of guidance services: parent, beginning graduate student, and the undergraduate student. For the latter category, Miller offers at the end of each chapter a study aid labeled "review and application." Throughout the book he suggests specific guidance activities for three levels of instruction: elementary, junior high school, and senior high school. Before offering these suggestions, Miller usually has a thorough review of the issues as well as a rationale that underlies them. He uses reference material from a variety of sources. However, the reviewer would challenge a statement made on the advertising flier that accompanied the textbook that "most of the publications cited are less than five years old." In summary, one

could say that this is a well written book that is comprehensive in scope and replete with examples.

HENRY KACZKOWSKI
University of Illinois

Guidance: An Examination by Ralph L. Mosher, Richard F. Carle and Chris D. Kehas (Editors). New York: Harcourt, Brace & World, 1965. Pp. vi + 232. \$2.50.

This paper-bound book is a revision and expansion of the 1962 Special Issue of the *Harvard Educational Review*. Four additional essays were added to the original nine articles. The papers can be divided in two general classes: those that deal with counseling issues and those that deal with guidance issues. The latter category discusses the following matters: impact of the Progressive movement on guidance; assumptions and practices underlying guidance; role of the counselor; cultural influences; impact of mental hygiene on guidance; theory of purposeful action; role of the computer in guidance; and guidance in a university setting. The following counseling theories are discussed: behavioral, Rogerian, existential, and ego. In addition Gordon Allport has an essay on the psychological models of man and on their guidance implications.

The aim of the editors was to present a "critical and scholarly review" of pertinent issues which were not concerned with the common stereotypes of guidance. Thus, they concentrate on a search for theoretical models and on a presentation of substantiating evidence. These objectives were fully met. The essays, which are lucid, adhere to the point at hand. Surprisingly, some of the theoretical models can be made operational in a school setting. The counseling papers not only serve as a springboard for review of existing theoretical positions but also contain additional refinements. They too contain suggestions as to how they can be implemented in a school setting.

As with any textbook, one must ask himself the question of possible use in the classroom. The reviewer found that the original articles can be used effectively in a theory of counseling course. The usefulness lies in the fact that they do discuss operational procedures. Although the text could be employed in a seminar course where issues are discussed, one would hesitate to recommend its use in an introductory guidance course. The typical graduate student lacks the background to read the book with profit. It takes a degree of sophistication to comprehend Allport's argument against man as a reactive being or Michael and Meyerson's use of reinforcement theory in counseling. The reviewer found that the original articles served as an excellent summary in a theory of counseling course. The four new papers will be an additional aide. Subsequent to a discussion of a myriad of theories, the papers help to point out some

of the administrative ramification of the many facets of the counseling process. Prior courses serve as background and provide the needed incubation period of ideas so that issues can be examined with meaning.

This book should be read by everyone who is going through a counselor preparation program. Those who are not in the area of guidance can use it to obtain a scholarly review of pertinent issues.

HENRY KACZKOWSKI
University of Illinois

Organization and Conduct of Guidance Services by Lester D. Crow and Alice Crow New York: David McKay, Inc., 1965. Pp. xi + 692.

"Example rather than precept" is the guideline used by the authors in writing this book. To many graduate students in counselor preparation programs, directors of guidance, and principals who are interested in specific guidance activities at various levels of instruction, this book will become a reference manual. On the other hand, the counselor-educator may not show the same degree of enthusiasm. He may even go so far as to label it a "non-book": a collection of activities that have been done in a variety of school settings.

The basic issue at hand is "What should a textbook in a specialized field contain?" The book that has a heavy emphasis on theory tends to be spurned by the practicing educator, for he is concerned with operational matters. When assigned to read a theoretically oriented book, the practitioner will badger the instructor with statements like "This may be fine on paper, but you take my school. . ." Typically, a compromise is effected so that a textbook will contain elements of theory and practice. The practitioner is still not placated, for he can rarely match his school's concerns with the one or two illustrations found in the text. The net result is a publishing dilemma: the instructor orders the book, but the students buy them. Which audience do you serve? Apparently Crow and Crow are of the opinion that a comprehensive presentation of specific activities enhances the learning process.

Each chapter of the book has a generalized introduction to the area under consideration. This is made up of a short synthesis of general issues, principles, assumptions, purposes, and goals, peculiar to the area. Each of these points may or may not be illustrated by example from actual school situations. The next step is to reduce the "guidance service" into a series of activities normally classified under this heading. Each heading in turn has a short generalized introduction in form of a synthesis of the rationale used to justify its operation. Theoretical points are illustrated by showing how they operate in specific situations. There is no attempt on the part of Crow and Crow to explicate a given point of view in guidance. Instead, they demonstrate how the more commonly held notions of

guidance are made operational. For example, if an individual is interested in knowing how a cumulative record is used on the elementary level he turns to the appropriate section. If he has a question about how a counselor can use the cumulative record, he can find an answer to it. However, the example may not necessarily be the same as the one previously used.

The book is divided into five parts: Introduction; Organization and Personnel in Guidance; Guidance in Action; Evaluation in Guidance Services; and Organization of Guidance in State and School Systems. It is further subdivided into 20 chapters. The section on "personnel" discusses the roles of the administrator, counselor, teacher, and specialist, as well as physical and budgetary matters. The section on "guidance in action" has a typical service orientation: inventory, information, counseling, placement, follow-up, and group. The section on "organization" discusses organization procedures on various levels of instruction: state, county, district, cities between 30,000 to 80,000 population, and large cities.

The reviewer has some difficulty coming to a definite conclusion about the merits of this book. There is no doubt in his mind that it can have a high degree of utility in certain educational settings. Educational practitioners will find it useful when they want to know how other school systems have carried out a given guidance activity. For example, what have other schools included in their list of objectives of guidance? Most counselor educators shudder at the thought that a guidance program should be developed by this comparison method. On the other hand, when the counselor-educator wishes to illustrate a point, the book offers an example or two on almost any guidance topic. However, the book has only a cursory discussion of theoretical matters. As a consequence, whether this book can be labeled "good" or "bad" depends on the criteria used. If one uses the concepts of "structure and procedure," then the book is very good; but if one employs the concept of "theory," the book is poor.

HENRY KACZKOWSKI
University of Illinois

Personality and Adjustment by Richard S. Lazarus. Englewood Cliffs, New Jersey: Prentice-Hall, 1963. Pp. x + 118. \$1.95 (paperback), \$3.95 (cloth).

Clinical Psychology by Julian B. Rotter. Englewood Cliffs, New Jersey: Prentice-Hall, 1964. Pp. xv + 112. \$1.95 (paperback), \$3.95 (cloth).

Both of these titles, along with a dozen others, are members of the Foundations of Modern Psychology Series. The series has an ambitious goal for itself: the replacement of the usual within-one-cover text for a course in introductory psychology by a collection of titles that individually covers the chapters of the text. One can hardly quarrel with the basic premise that psychology in the sixties

is too complex for one author to do justice to all chapters with equal expertise. The publisher's solution is to have each section written by an acknowledged authority and participant, but published separately to allow the instructor freedom in selecting his chapters.

One can imagine each contributor receiving his instructions to write such a section in no more than 120 printed pages, and for the intended audience of students about to begin their first course in psychology. And no peeping to see what the other contributors are doing. The results are destined to turn out as they appear to be when these two books are compared. Rotter's contribution makes no assumption of sophistication on the part of the reader and provides a rather readable overview of clinical psychology. The focus is on the activities of the clinical psychologist. Theory, except in the discussion of psychotherapy, is ignored. A glossary of terms is provided to help the student grappling with this material for the first time. Rotter, incidently, seems to be the only contributor in the Series who thought of providing such an aid for the beginning student.

Lazarus, who is also the editor of the Foundations of Modern Psychology Series, handles an area which necessarily overlaps that of Rotter. While one might read Rotter to find out what clinical psychologists *do*, one would have to read Lazarus to find out what clinical psychologists *think*. His approach is to emphasize theory (S-R, Lewinian, psychoanalytic, Rogerian) and research findings as he deals with the structure, development, and assessment of personality. The difference in two approaches to the handling of the assignment perhaps can best be summarized by stating that Rotter's contribution is for the introductory student only, while that of Lazarus could be useful with courses in personality. Other titles in the Series (*cf* Hochberg's *Perception*) would provide some rough going for the beginning student looking for help in his text.

Of the two books under consideration, the reviewer has a marked preference for that of Lazarus. This may be because Rotter comes nearer to the goal of presenting material for the beginning student, while Lazarus' contribution is valuable at several levels in the educational process.

What is the place of the Series in the college or university bookstore? It is difficult to believe that many instructors will discard their hardcovered Morgan, or Munn, or Hilgard, or whatever, to be replaced by selected books from the Series, at \$1.95 per chapter, to form a collection of paperbacks. The logical next move by Prentice-Hall is to combine the entire series in a single hardcover edition, like Hilgard, or Morgan, or Munn.

PHILIP HIMELSTEIN
Texas Western College

Influencing Through Argument, by Robert B. Huber. New York: David McKay Co., Inc. 1963. Pp. vii + 392. \$5.00

Professor Huber has taught at Manchester College and at the Universities of Indiana, Wisconsin, Oregon, and Vermont. As of the writing of *Influencing Through Argument*, he is Chairman of the Department of Speech at Vermont. The book was written because, "For several years, the author taught classes in argumentation only to meet with the discouragement of having the students fail to apply the test for various forms of reasoning, and the revelation of various types of fallacies in the actual speaking situation." It is undoubtedly a sad commentary on psychologists' failure to communicate principles of learning and transfer to colleagues in other disciplines that Professor Huber tried to solve this problem by writing a book.

The book is somewhat confused as to purpose. One purpose is to provide a general treatment of verbal persuasion from the point of view of the persuader. For this purpose, the book is substantially weakened by a complete failure to consider scientific investigations of persuasion conducted by psychologists and others. The absence of references to the work of the Hovland group is particularly conspicuous. A second purpose, and the one that actually seems dominant, is to provide a guide to winning intercollegiate debate tournaments. Since the author has served for many years as a judge of debate tournaments, a number of valuable hints relevant to this purpose are provided.

The author consciously relates his book to the classical tradition of Western thought, and in particular to Aristotle's *Rhetoric*. This orientation was quite influential in determining the structure of *Influencing Through Argument*. The book begins with two chapters justifying the use of argument to influence others and discussing with what kinds of subjects argument is likely to be effective. This treatment is followed by two chapters on the analysis and gathering of material for arguments, and five chapters on influencing through evidence, through induction, through deduction, through causal reasoning, and through reasoning from analogy. This section is followed by one chapter on refutation of the arguments of others. In this chapter, the emphasis on the debate tournament context is especially clear. The book concludes with four chapters on various aspects of writing and delivering speeches.

Each chapter is carefully outlined. This outlining has the strong advantage of making the author's intentions very clear. However, it also makes the many errors of logic and the many confusions of thought very clear. The over-all impression is that the author does not really understand what he is writing about. In order not to belabor this point, only two examples will be given, both taken from the chapter on deduction. First, the fallacy of lack of dis-

tribution of terms involves the distinction between "some" and "all." It does not involve, as the author's examples suggest, omitting both "some" and "all" from the syllogism. Second, the following can hardly be called a syllogism at all, and so is not an example of lack of distribution of terms.

"Vermont is a Republican state.

John Jackson is a Vermonter.

Therefore, John Jackson is probably a Republican."

In addition to these specific examples, the author continually confuses the truth of the major and minor premise with the rules of deduction.

Is there then any reason why a reader of this journal might want to purchase this book? In spite of its defects, it does provide a guide to writing and delivering speeches which could be helpful. For this purpose, however, many better sources are available. The reviewer feels, therefore, that readers of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT will have nothing to gain from purchasing this book.

JAMES M. RICHARDS, JR.
American College Testing Program
Iowa City, Iowa

ERRATUM

An error has been noted in the article entitled "Intellective Predictors of Success in Nursing School" by Jon M. Plapp, George Psathas, and Daniel V. Caputo, which appeared in Volume XXV (1965), 565-577. The test referred to as the *Scholastic Aptitude Test (SAT)* of the College Entrance Examination was not this test, but was the *Nursing Admissions Test (NAT)* of the Scholastic Testing Service. All references to the *Scholastic Aptitude Test* or its abbreviation, *SAT*, should therefore read as *Nursing Aptitude Test* or its abbreviation, *NAT*.

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of October 23, 1962; Section 4369, Title 39, United States Code)

1. DATE OF FILING
September 23, 1965
2. TITLE OF PUBLICATION
Educational and Psychological Measurement
3. FREQUENCY OF ISSUE
Quarterly
4. LOCATION OF KNOWN OFFICE OF PUBLICATION (Street, city, county, state zip code)
2901 Byrdhill Road, Richmond, Virginia 23228
5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS (Not printers)
3121 Cheek Road, Durham, N. C. 27703
6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR
PUBLISHER (Name and address)
G. Frederic Kuder, Box 6907 College Station, Durham, N. C. 27708
EDITOR (Name and address)
G. Frederic Kuder, Box 6907 College Station, Durham, N. C. 27708
MANAGING EDITOR (Name and address)
Geraldine R. Thomas, 3121 Cheek Road, Durham, N. C. 27703
7. OWNER (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)

NAME	ADDRESS
G. Frederic Kuder (Owner)	Box 6907 College Station, Durham, N. C. 27708
8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES (If there are none, so state)
None
9. THIS ITEM MUST BE COMPLETED FOR ALL PUBLICATIONS EXCEPT THOSE WHICH DO NOT CARRY ADVERTISING OTHER THAN THE PUBLISHER'S OWN AND WHICH ARE NAMED IN SECTIONS 132.231, 132.232, AND 132.233, POSTAL MANUAL (Sections 4355a, 4355b, and 4356 of Title 39, United States Code)

I certify that the statements made by me above are correct and complete.

(Signature of editor, publisher, business manager, or owner)
Geraldine R. Thomas, Managing Editor

17 MAR 1972



17 MAR 1972